THE UNIVERSITY
of EDINBURGH

Quantitative Fluorescence Microscopy Methods for Studying Transcription

with application to the yeast GAL1 promoter

Elco Bakker

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

to the

University of Edinburgh

THE UNIVERSITY
of EDINBURGH

September 2015

quantitative fluorescence microscopy methods for studying transcription, with application to the yeast GAL1 promoter

Elco Bakker

April 15, 2016

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise. The work has not been submitted for any other degree or professional qualification.

Chapter 3 was collaborative work undertaken by various lab members, and in particular M Crane, who wrote the original programmatic structure, the algorithm for cell centre identification by support vector machine and for tracking. I have attempted to assign credit clearly throughout the chapter.

Chapter 5 was jointly undertaken with Lucia Bandiera, a visiting student from the University of Bologna. Though I envisioned and guided the project, experiments and analysis were carried out together.

Elco Bakker

Abstract

The advent and establishment of systems biology has cemented the idea that real understanding of biological systems requires quantitative models, that can be integrated to provide a complete description of the cell and its complexities. At the same time, synthetic biology attempts to leverage such quantitative models to efficiently engineer novel, predictable behaviour in biological systems. Together, these advances indicate that the future understanding and application of biology will require the ability to create, parameterise and discriminate between quantitative models of cellular processes in a rigorous and statistically sound manner. In this thesis we take the regulation of *GAL1* expression in *Saccharomyces cerevisiae* as a test case and look at all aspects of this process: from reporter selection to data acquisition and statistical analysis.

In chapter B we will discuss optimal fluorescent reporter selection and construction for investigating transcriptional dynamics, as well as procedures for quantifying and correcting the various sources of error in our microscope system.

In chapter 3 we will describe software developed to analyse fluorescent microscopy images and convert them to gene expression data. A number of iterations of the software are tested against manually curated data sets, and the measurement error produced by its imperfections is quantified and discussed.

In chapter 4 a method, based on fluctuations in photobleaching, is developed for quantifying both measurement error and the relationship between protein con-

centration and measured fluorescence. The method is refined and its efficacy discussed.

In the last section I make a preliminary application of these methods to investigating the regulatory effect of the *GAL10*-lncRNA. Interesting phenomena are observed and further investigated using two new strains: genetic variants expressing a fluorescent reporter from the *GAL1* promoter, one harbouring a wild type *GAL1* promoter and one in which the binding site for the *Gal10* noncoding RNA has been removed. The methods developed throughout the thesis are applied and the data analysed. A heterogeneous response, distinguishable between the two strains, is observed and related to cell-to-cell variations in growth rate.

Lay Summary

It is clear to anyone inhabiting this world that living organisms are capable of producing diverse behaviours and complex structures ranging from the microscopic to the macroscopic; that they are adaptive, self replicating, and far more sophisticated that any technology we produce. That you are reading and comprehending this document, while the computer in front of you can only display it, must be evidence of that. It is also part of the common conciousness that this complexity and sophistication is somehow encoded in an organisms DNA, but our knowledge of the mechanisms by which DNA determines the structure and behaviour of living things is still far from complete.

What is known is that certain DNA sequences encode proteins, that molecular machinery 'reads' this sequence to synthesises or 'express' the protein, and that the proteins present in a cell largely determine its behaviour. As an example, in multicellular organisms like mammals it is the expression of a particular collection of proteins over the cells lifetime that determine its cell type - turning one cell into a muscle fibre and another into a neuron - even though they both share the same DNA. Clearly the control of protein expression is immensely important, and it is thought that much of the DNA that doesn't encode proteins is devoted to this process of gene regulation.

Gene regulation has been studied for over 60 years, which has revealed it to be a dauntingly complex process involving many different reactions between a panoply

of cellular constituents. Further, it has been found that even between genetically identical cells exposed to the same conditions the protein expression can vary dramatically: a consequence of the small size of cells and the inherently random nature of chemical reactions, which becomes apparent when small numbers of molecules are involved. These two properties - complexity and randomness - make it difficult to understand the behaviour of biological systems even if the underlying rules that govern the system are understood. An analogy could be drawn with a radio. If you were to open one, identify the components and look up their function, you might still find it very hard to understand or predict its behaviour - and that is a completely predictable system designed by another human being. No conciousness has every had to understand the workings of a cell. In contrast to human inventions its evolution has been guided only by chance and function, not by any need for comprehensibility, and as such its complexity can far exceed those of human inventions.

Faced with such a problem the only way to proceed is to attempt to develop rigorous mathematical descriptions of the component parts, and combine them into a model of the cell that can explain and predict behaviour in a way that is not possible by simple inspection. As part of this effort, we have in this thesis developed experimental tools that allow us to construct and test such mathematical descriptions.

We focus on the bakers yeast *Saccharomyces cerevisiae* . Though this may seem an irrelevant subject, it shares many proteins and regulation mechanisms with humans and as a 'simple case' of a biological system has provided many insights applicable across the species. In collaboration with other members of our research group, we have developed computational methods to automate acquisition of protein expression data from thousands of individual yeast cells simultaneously observed over their entire lifetimes. Not only that, we can also extract secondary information such as their size, shape and the cellular division events: enriching

our data with contextual information. This capacity provides a means to understand the regulation of protein expression in incredible detail.

We apply these tools to a system regulated by a long non-coding RNA - a subtle and ubiquitous regulatory element - and find that we can reveal new insights into the variation in gene expression between cells and the connections between gene regulation and cellular behaviour. Beyond this, we make a first foray into the difficult problem of using this data to build mathematical models of gene expression.

Acknowledgements

It is rare to be given the opportunity to thank in publication the people that have contributed to a period of ones life, and I will take that opportunity here.

I would first like to thank my two supervisors, Peter and Ian. Peter for providing support, guidance and insight throughout this project. He is an exemplary scientist, and there are few students outside of our lab who receive as much dedicated attention and quiet co-contemplation from their supervisors as we do. Ian I would like to thank for four years of near paternal concern. Providing advice in all matters of laboratory and biology, but more than that, for encouragement and reassurance throughout the PhD. In this vein of mentorship, I would also like thank Alan Cordeaux, John Chalker and Detlef Dürr: great teachers all.

I would like to give my thanks to all the swain lab members: past, present, permanent and transient. Whether discussing our work, unpicking each others code, eradicating each others armies from Westeros or the Koprulu Sector or pursuing rogue magi through city streets and slums: you have contributed immeasurably to my enjoyment of the last four years. To two member I would like to give special thanks. Firstly, to Lucia Bandiera: the most industrious Erasmus student ever known who played and unwavering midwife to my final chapter. Secondly

to Matt Crane. Not only for your work in the lab, on which all mine relies, but for enthusiasm, discussions, support and guidance. For squash, walks, Swann breakfasts and an understanding ear when I needed it. Students in the Crane lab will be the envy of their peers.

Last I would like to thank my family: I promise that after this I really will get a job stop being a student, at least for a time.

Contents

List of Figures

List of Abbreviations

ALCATRAS	A Long-term Culturing And TRApping System
UAS	upstream activation sequence
Pol II	RNA polymerase II
PIC	preinitiation complex
RSC	chromatin remodelling complex
HAT	histone acetyltransferase
HDAC	histone deacetylase
FISH	fluorescence in situ hybridisation
TSA	Trichostatin A
HOG	High Osmolarity Glycerol
DIC	Differential Interference Contrast
ODE	ordinary differential equation
EMCCD	lectron multiplication charge coupled device
PSF	point spread function
PDMS	Polydimethylsiloxane
PFS	perfect focus system
GFP	green fluorescent protein
CFP	cyan fluorescent protein
YFP	yellow fluorescent protein
EGFP	enhanced green fluorescent protein
SC(media)	synthetic complete

SVM	Support Vector Machine
GUI	grahical user interface
MCMC	Markov chain Monte Carlo
FCS	Fluctuation correlation spectroscopy
DPP	dynamic prior propagation

Chapter 1

Introduction

Transcriptional regulation, the process by which a cell controls the amount of a particular mRNA that is transcribed, has been studied for over fifty years and is important to many areas of medicine and biology. Transcription misregulation can cause a wide range of diseases and disorders, and in development cell fate is largely determined by the transcriptional program in the cell. In the nascent field of synthetic biology transcriptional regulation has a foremost role, underpinning many of the genetic circuits constructed.

I will begin with a short summary of the general picture of transcriptional control in *Saccharomyces cerevisiae*; for more information please see recent reviews [70, 143]. We will confine the discussion to genes transcribed by RNA polymerase II (Pol II), since these are the ones we will investigate in this research project.

1.1 Transcriptional Regulation in *Saccharomyces cerevisiae*

Transcription initiation in yeast is a complex process involving a large number of cellular species. It is largely controlled by a regulatory region 5' of the open reading frame which is divided into the promoter, which often contains a TATA box, and upstream activation sequences (UASs).

Before transcription can begin a preinitiation complex (PIC), consisting of a number of general transcription factors (proteins and protein complexes that are necessary for transcription), must form at the gene promoter. Preinitiation complex formation is followed by the binding of RNA polymerase II (Pol II) and other general transcription factors. Pol II then undergoes a number of conformational changes which cause it to begin elongation. At this point transcription initiation has been completed.

Transcription initiation will not generally occur autonomously but requires certain co-activators, making these co-activators a common end point for transcriptional control mechanisms. The two most important co-activators for preinitiation complex formation are TFIID and SAGA. Though both these complexes have the general transcription factor TBP (TATA Binding Protein) as a subunit, they control distinct sets of genes. Genes controlled by TFIID are generally labelled 'growth genes': they are fairly constitutive in their expression, have no TATA box in their promoter and are relatively noiseless. Genes controlled by SAGA are labelled 'stress genes': they display large changes in expression in response to extra cellular conditions, have a TATA box in their promoter and are more noisy [7]. Approximately 90% of Pol II transcribed genes are growth genes, with stress genes making up the other 10%. The third important co-activator is mediator. This complex binds Pol II, providing an intermediary by which Pol II can be recruited, and also stabilises the PIC.

The proteins that control the rate of transcription of particular genes are called transcription factors, and in yeast the majority work by recruitment: the simple process of localising some appropriate species to the promoter of the gene in question. In the case of transcriptional activators it is usually either one of the co-activators mentioned above or a chromatin modifying enzyme (see discussion below) that is recruited. Transcriptional repressors will generally recruit complexes that maintain chromatin in a repressive state. The activity of these transcription factors is often controlled by their concentration, but can also be controlled by nuclear localisation or by modifications, such as phosphorylation, which activate or de-activate the transcription factor.

1.1.1 The role of Chromatin in Regulation

As will already be clear, chromatin is an important component of transcription regulation in yeast. The majority of chromatin based transcriptional control seems to work by gross occlusion: DNA wrapped around a nucleosome is inaccessible to Pol II and the general transcription factors, and is therefore not transcribed. For this reason the position of nucleosomes on the DNA is carefully controlled. Here again we see a difference between growth and stress genes. While growth genes maintain a nucleosome free region around their promoters, stress genes display a high density of nucleosomes over theirs. Transcription of stress genes therefore requires the eviction of nucleosomes from the promoter, which may contribute to their noisy expression[144]. Eviction is usually effected by the recruitment of chromatin remodelling factors. Important examples are Swi/Snf, Chromatin Remodelling Complex (RSC) and Isw1,2. Of these RSC is the most significant, regulating hundreds of genes and maintaining the nucleosome free region of many growth genes, while Swi/Snf regulates a smaller set of stress genes

and mostly in conjunction with other regulators such as SAGA. *isw1/2* mutants shower milder phenotypes than either of the other chromatin remodelling factors, but Isw1/2 have the interesting property of sliding nucleosome laterally rather than evicting them, and are shown to have detectable effects on the chromatin of approximately 400 genes [190].

Chromatin structure is important for transcription initiation, but transcription also affects this structure. High rates of transcription increase the size of the nucleosome free region over the promoter and cause denser nucleosome occupancy in coding regions, and very high rates of transcription actually cause sparser occupancy of the coding region. Much of these changes have been attributed to the effects of Pol II transit and *in vitro* studies have shown that Pol II leaves behind a trail of 'damaged' nucleosomes (nucleosome missing some of their components) which are easily removed if the gene is transcribed again.

Histones removal by any of these processes requires histone chaperones: complexes that bind free histones and escort them to and from the DNA. These species are therefore important in the processes described above and have been implicated in the timing of gene activation as well as the maintenance of silenced regions of DNA.

Another aspect of chromatin mediated gene regulation is chromatin modification. Though the core of the histone is fairly inaccessible each histone has an accessible tail which can be modified to affect the histone's behaviour. Although there are a broad range of modifications that can be made, which have diverse effects on transcription, it is broadly the case that acetylation by histone acetyltranferases (HATs) make a region more competent for transcription, whereas histone deacetylases (HDACs) cause the DNA to be less accessible and therefore repressed. A number of HAT containing complexes, such as SAGA and NuA4, are important for transcriptional control and also elongation, probably via regulation of RSC complex activity. Histone modification is also affected by transcription, and this

effect is used as a transcriptional control method in cases like *GAL1-10*. Here, the transcription of a non-coding RNA that traverses the *GAL1-10* gene cluster causes deacetylation of the coding region, repressing transcription of the genes [84].

With contributions from so many interconnected processes it seems unlikely that human inspection alone will avail in providing a satisfactory description of transcriptional regulation. Even if it did, the result could only provide limited assistance in the design of synthetic circuits and could not be integrated into larger, systems level, models of cells and cellular processes. For these purposes, and for a complete understanding of transcriptional regulation, a more rigorous quantitative approach will be necessary. This consists at its most basic of constructing a model capable of quantitative predictions and comparing these predictions with data: phenomenological or quantitative. We will now undertake a brief description of the type of models we will be interested in for describing transcriptional control before reviewing efforts to apply such models to biological systems and in particular eukaryotes.

1.2 Modelling Transcription

The models we are interested in are those based on the chemical master equation. Although more involved models of transcription, such as those based on molecular simulations [170], have been undertaken these are better suited to understanding the conformation of proteins and complexes during DNA binding and transcription than understanding the cell level process of transcriptional control. The chemical master equation is a high level description of transcription in which the species of interest are assumed to be well mixed within the cell, or at least

within each subcellular compartment. The promoter is described as being in one of a number of discrete states, each with its own transcription rate. Transcription events are uncorrelated given a particular state of the promoter and, like all reactions in a chemical master equation description, are assumed to happen instantaneously. If there is only one state of the promoter then the transcription rate is constant and, provided the decay rate of the mRNA is also constant and first order, the mRNA content of individual cells will be described by a Poisson distribution. If there are numerous states of the promoter than there is the possibility of transcriptional regulation. If, for example, a possible state of the promoter is one with a transcription factor bound then the concentration of this transcription factor will determine the proportion of time the promoter spends in the 'active' transcription factor bound state. We will in general refer to these models - master equation descriptions of transcription in which the promoter can be in one of a discrete set of states with differing transcription rates - as promoter state models. The chemical master equation can easily accommodate translation, interactions of proteins, and networks of genes, though the increased complexity makes simulation and inference time consuming: often prohibitively so. This can to some degree be overcome by applying various analytical approximations available for the master equation [193], but these are in general only accurate when the number of molecules involved is large or reactions can be clearly partitioned by time scale. Even in these cases the approximations can produce inaccurate or misleading results, especially if the system considered has feedbacks [152].

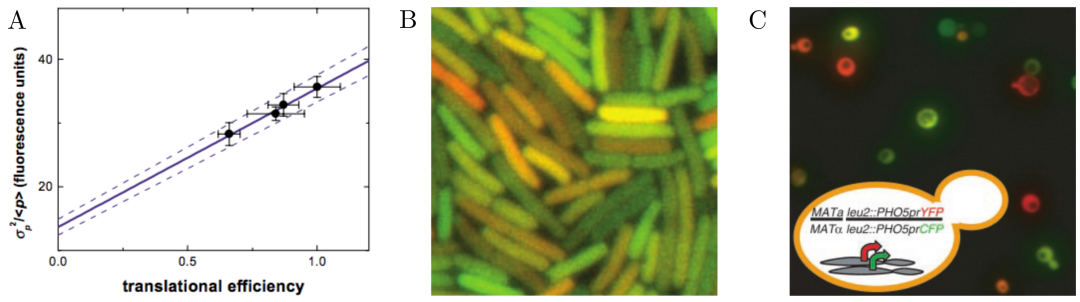

Figure 1.1: Landmark papers from the inception of quantitative, stochastic analysis of gene regulation. Panel A shows the relationship between translational efficiency and phenotypic noise strength measured by Ozbudak et al. [131]. This was one of the results which established that gene expression could be understood in terms of the simple, master equation models described. Panels B and C show images from Elowitz et al. [46] (*E. coli*) and Raser and O'Shea [144] (*Saccharomyces cerevisiae*) respectively. Both are overlay images from two colour experiments; the difference in colour reflect intrinsic noise, while the difference in brightness is indicative of extrinsic noise. All are reproduced with permission.

1.3 Applications of Models of Transcription to Biological Data

1.3.1 Analysis of Population Variation

Over the last fifteen years models of the sort described have been applied to diverse systems and data types, particularly at the single cell level where stochastic variation between cells makes intuition even more difficult. Some of the earliest work was by Ozbudak et al. [131], in which they use flow cytometry analysis of *Bacillus subtillis* to show that the degree of cell to cell variation varies across genes and depends on the biochemical parameters of those genes. This was important in establishing that the type of model described above could be applied to transcription at all, and that the variation they predicted was not in general dominated by other cellular processes. In the same year Elowitz et al. [46] used an innovative two colour method in *E. coli* to distinguish intrinsic variation, the variation between cells and over time due to the stochastic nature of the mod-

els described above, and extrinsic noise, the variation between cells that remains after this intrinsic variation has been accounted for. This work showed that, especially in the high expression regime, extrinsic noise was dominant. This somewhat contradicts the work of Ozbudak et al. [131], in that it shows that a significant component of cell behaviour cannot be explained by the simple promoter state models of the type described above. Later work in yeast by Raser and O'Shea [144] showed that extrinsic noise could also dominate in eukaryotes. Focusing on the intrinsic noise the authors were able to discriminate between different two state promoter models for gene regulation and to some degree corroborate these findings with studies of mutated promoters. These papers typify the paradigm for the quantitative study of single cell data: cell to cell variation could be partitioned into intrinsic and extrinsic components, 'two colour' experiments could distinguish between these two sources and intrinsic noise could provide quantitative information about the underlying system not accessible by measurements of the population average alone. Figures from these landmark papers are reproduced in figure 1.1.

Over the next ten years numerous papers used fluorescent reporters to measure intrinsic variability and infer models of biological systems. Large scale studies of the cell to cell variation of many proteins indicated that genes were expressed in infrequent transcriptional bursts, and that the degree of variation was dependent on the controlling transcription factor [126]. Careful analysis of flow cytometry data was used to parameterise quantitative models of operons in *E. coli* [123] and promoters in yeast [25, 128] and mammalian cells [158]. Later, quantitative noise analysis of yeast histone deacetylase (HDAC) heterozygotes [189] and yeast cells harbouring mutated promoters [83] contributed to understanding the genetic and biochemical underpinnings of promoter state models. Broadly, the presence and condition of the TATA box was determining of the transcription rate from a permissive state [83], while different HDAC operated in different ways: some

modulating transcription rate from the permissive state and some regulating its duration [189].

In the same period mRNA visualisation in single cells by fluorescence in situ hybridisation (FISH) was used to measure the distribution of mRNAs in a population. This work [142] showed that in mammalian cells genes were also expressed in transcriptional bursts, indicative of multistate promoter models. Bursts were correlated for genes close to each other on the genome, a result also observed using fluorescent proteins in yeast [8]. Quantitative analysis of FISH data showed that one and two state promoter models were appropriate for different genes, depending on the transcription factors which regulate them [196]. Later experiments in which two species of mRNA could be observed simultaneously showed that, while there was some extrinsic influence producing global correlations in mRNA expression [54], mRNA variation was dominated by intrinsic noise [81]. Like protein noise data, mRNA distributions obtained by FISH were used to infer models of gene expression, and to draw connections between the parameters and processes in these models and the biochemical knowledge available about the underlying regulation [125, 162].

Another type of experimental data analysed using a multistate description was nucleosome occupancy. Nucleosomes had quickly emerged as a possible mechanism for generating the multiple *cis* acting promoter states seen in population variation data [144], and observation and perturbation of nucleosomes provided an opportunity to confirm the biochemical nature of these putative states. Work by Brown and Boeger [20] confirmed that even closely positioned promoters had intrinsic fluctuations in their nucleosome occupancy state, and that these intrinsic fluctuations occurred even without transcription. In the same year Small et al. [161] showed that cells sorted into expressing and non expressing populations also showed distinct nucleosome configurations in their promoter regions, with expressing cells having a large nucleosome free region in their promoter.

In particularly interesting work the year before, Brown et al. [21] used electron microscopy to measure the population distribution of different nucleosome states in the *PHO5* promoter. By comparing the occupation probabilities with FISH estimates of the proportion of transcriptionally active cells, they attempted to partition nucleosome states as transcriptionally active or inactive and provide a biologically reasonable interpretation.

While this is an extremely impressive effort to give a biochemical interpretation to the abstract concept of promoter state, it can not be taken as a general conclusion that promoter states are related to nucleosome occupancy. Despite the absence of nucleosomes, bacteria also display distinct transcriptional states [62, 162], which recent work ascribes to DNA supercoiling [31], and the general cause of distinct promoter states is still unknown [76].

1.3.2 The Source and Nature of Extrinsic Noise

There has also been great interest in the biological source and implications of extrinsic noise. Work by Rosenfeld et al. [148] in *E. coli* showed that extrinsic variations in the gene regulation function (the putative function relating transcription rate and transcription factor concentration) had autocorrelation times of around a cell cycle, and thereby implicated stable cellular constituents as contributors to extrinsic noise. By labelling genes in the same pathway in yeast with distinguishable fluorescent proteins, Colman-Lerner et al. [35] were able to show that both variations in upstream regulators and in cell cycle stage contributed to extrinsic noise; conclusions that were later found to be true on a larger scale [167, 180]. Detailed time series analysis of yeast cells harbouring both a fluorescent cell cycle reporter and distinguishable alleles of the same gene showed an increased coordinated transcription at S/G2 phase [199]. Work in *E. coli* indicated a role for global physiological state [13], while recent experiments in

mammalian cells points to some global control to maintain transcript concentration at a fixed level [97, 132]. Being global effects, these can all contribute to extrinsic variability depending on the system observed and the experimental methodology employed.

To summarise, the body of work described shows that a great deal has been learnt from the quantitative study of gene expression and variation in single cells. Simple multi-state promoter models are often effective at describing experimental data, with the single state and two state models being appropriate for constitutive and regulated genes respectively. Single cell data of fluorescence and mRNA can be used to infer parameters of such models, and often indicate that transcription factors control the transition rate between transcribing and non transcribing states, leading to 'bursty' mRNA expression. In many experiments the biochemical nature of these promoter states seems to be related to promoter nucleosome structure, but this is not universally the case. While these simple models can explain much of the single cell protein data there is also a significant contribution from extrinsic factors such as cell cycle stage, upstream transcription factor concentration and global physiological state [95, 122, 141].

1.3.3 Single Cell Time Series Data

While the work above is impressive, it largely relies on measurements of population distributions rather than single cells over time. Even when the population distribution is evolving in time, these measurement are limited in their ability to discriminate between models of gene expression and to understand the temporal evolution of extrinsic noise. Single cell time series is richer [109], and has been

used to address some of these questions over the last ten years. Early time series analysis from Rosenfeld et al. [148] in 2005 showed that extrinsic variation between cells had an autocorrelation time of around a cell division. Work from the same year by Golding et al. [62] used time series data of mRNA content of single *E. coli* cells to provide extremely strong evidence for a two state model of gene expression.

In 2011, two papers [74, 172] used detailed statistical analysis of fluorescence time series from mammalian cells to show that a two state model of gene expression was not consistent with their data, and that a third - refractive - state was required. Use of two distinguishable reporters in Harper et al. [74] showed that these states were intrinsic to the gene, and perturbations with Trichostatin A (TSA, a histone deacetylase inhibitor) in both papers indicated a role for nucleosomes in the establishment of these distinct promoter states. Later work by Molina et al. [120] on experimentally stimulated mammalian cells showed that analysis based on the two state model could still be informative, even in cases where the two state model is not applicable, and applied such an analysis to reveal detailed quantitative information about the temporal transcriptional response of cells to naturally occurring stimuli. Though difficult to generalise, this is an important observation, since we know from our detailed biochemical knowledge that the two state model of transcription is a simplified description. Work from the same year by Finkenstädt et al. [51], on cells challenged by translation and transcription inhibiting drugs, showed how analysis of dynamical observations from single cells could be used to infer parameters on a cell by cell basis. Providing a measurement, albeit indirect, of the extrinsic variability in model parameters across a population of cells.

The work described illustrates the understanding that can be obtained from single cell time series data. As shown by Molina et al. [120] and Finkenstädt et al. [51], this is particularly true when the cellular conditions can be changed. n this vein,

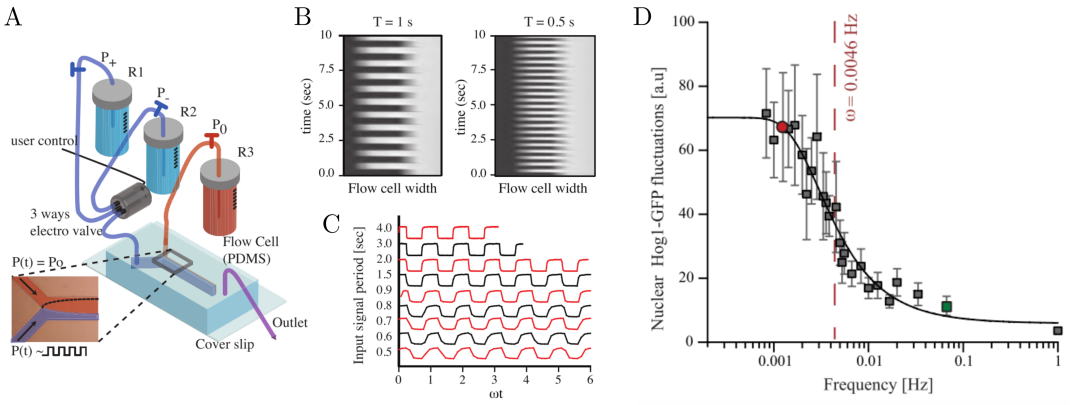

Figure 1.2: Results and microfluidic device from Hersen et al. [79]. Panel A shows the structure of the microfluidic device. A 'y-channel' accommodates the cells and allows the rapid switching between high and low osmolarity media; this is shown by the media profiles in Panels B and C. panel D shows the novel phenotype investigated: bandwidth. The x axis shows the switching frequency of the media, while the y axis shows the fluctuations in nuclear localisation of Hog1p. This phenotype, only accessible via microfluidics and microscopy, provides new insights into the behaviour and network properties of the well studied HOG pathway. Reproduced with permission from Hersen et al. [79].

we will next discuss microfluidics, a technology that facilitates far more complex dynamic environments.

1.4 Microfluidics and Dynamic Environments

Microfluidic devices, and PDMS microfluidic devices in particular [115] , provide a cost effective way to image cells while subjecting them to dynamic media conditions that are still more controlled than those in batch culture [47]. These devices generally use a shaped PDMS structure to hold cells against the surface of a cover slip, either by vertical restriction [10, 103] or by traps or jails [9, 39, 150]. Use of this technology has led to many interesting discoveries[9, 129], such as the observation of resilience to antibiotic challenge emerging from phenotypic variation [6, 182] and the importance of lineage effects in cell fate decisions [96].

The complexity of dynamic single cell data makes computational modelling a

natural complement to microfluidic experiments, and there have been a number of studies combining these two. Bennett et al. [10] used a microfluidic device employing vertical restriction to trap *Saccharomyces cerevisiae* cells expressing a fluorescently tagged Gal1 protein, and subject them to periodic activation with galactose and repression with glucose. By identifying discrepancies between their data and simulations of a literature based model of the galactose/glucose system, the authors were able to identify a glucose dependent change in the decay rate of certain GAL gene transcripts.

Ullman et al. [179] used a similar vertical constriction device to trap large colonies of *E. coli* . Leveraging the large single cell statistics so obtained, the authors were able to observe subtle, cell cycle dependent changes in the expression of LacI and LacY and expression correlations across lineages. Paliwal et al. [133] designed a microfluidic device for imaging budding yeast cells while exposing them to gradients of mating pheromone of various slopes. This allowed them to observe a bistable response in both schmooing behaviour and the expression of mating pheromone responsive genes. A computational model was used to investigate this bistability and suggest important network components that were interrogated by genetic manipulation. The use, as in this work, of modelling to suggest species responsible for an observed dynamic behaviour is an important complement to many microfluidic experiments, which are often too time consuming for large scale screens. This is to some degree overcome in two more recent papers on yeast mating factor response pathway from Carl Hansen's lab [50, 174]. Both papers use a similar device, with many isolated chambers that allow simultaneous imaging of a number of mutants in a range of media, either static [50] or dynamic [174]. Though these studies observe many interesting phenotypic consequences of the deletion of genes in the mating pheromone response pathway, the lack of computational analysis makes the data difficult to interpret.

Another system subject to intensive microfluidic study is the High Osmolarity

Glycerol (HOG) pathway in yeast, both because it is a model system for cell signalling and stress response and because one of yeast's major responses to osmotic stress is the nuclear localisation of Hog1p: a response most easily observed by microscopy. Two similar papers from 2008 both used dynamic inputs and engineering inspired computational analysis to investigate the signal processing of the HOG network. In Hersen et al. [79], a simple ordinary differential equation (ODE) model of the HOG network is used to calculate the bandwidth of the network: the amplitude of oscillations in Hog1p nuclear localisation as a function of the frequency of switching between high and low osmolarity media. To measure the bandwidth experimentally they construct a 'Y channel device', which allows rapid switching between two media and confirms the theoretically predicted form of the bandwidth. Investigating over-expression and gene deletion strains, the authors were able to draw connections between changes in the experimentally measured bandwidth and the network structure: furthering understanding of how the architecture of the network results in the dynamic behaviour observed. Figures describing the operation of the device and the bandwidth measurements are reproduced in figure 1.2.

Mettetal et al. [118] used a similar ODE model and microfluidic device that also allowed media to be quickly switched. In this case, the model was fitted to one set of data and corroborated on another. Frequency response was again observed and cycloheximide, a small molecule that inhibits translation, was used to distinguish the contribution of gene expression and post translational modification to the cells response.

Both these papers show how dynamic control of the media and single cell imaging can give access to a whole new phenotype, bandwidth in these cases, which is not accessible by any other means. Computational modelling allows these complex phenotypes to be understood in terms of the underlying biochemical network, and for that understanding to be tested using conventional genetic techniques.

The idea of frequency response has since been applied to understanding gene expression at the transcriptional level. Two recent papers from the O'shea lab [72, 73] used an exceptionally engineered yeast strain, in which nuclear localisation of the Msn2p transcription factor is controlled by concentration of the small molecule 1-NM-PP1 in the media. By using a microfluidic device to quickly switch the concentration of 1-NM-PP1, they are able to create transcription factor oscillations and observe the expression of downstream genes. The data generated is used to distinguish and parameterise a phenomenological three state promoter model for both synthetic [67] and naturally occurring [72] promoters. These studies, and those of Mettetal et al. [118] and Hersen et al. [79], show how dynamic media control and frequency response can be used to understand both networks and individual promoters.

While the work described in this section so far is technically sophisticated, in many respects the time series analysis is not as sophisticated as that applied by Suter et al. [172] or Harper et al. [74]. The models used are generally phenomenological ODE models that do not take account of cell to cell variation or stochastic fluctuations. Hansen and O'Shea [72] is in some ways an exception to this, as the authors both measure intrinsic noise and assess whether their promoter models are consistent with these measurement, but even here the models include unexplained phenomenological aspects and the noise measurements are not used for fitting but for verification after the fact. Only the work of Zechner et al. [195] combines both dynamic microfluidic data and rigorous statistical analysis. In their 2014 paper, the authors develop a novel statistical analysis method that allows efficient inference of master equation based models, with extrinsic variation in model parameters, from single cell traces of multiple cells. They apply this inference scheme to data from a synthetic system dynamically induced in a microfluidic device, and are able to parameterise and discriminate between various promoter state models with extrinsically varying translation rate. Having

selected the most likely model they are able to infer probable states for the unobserved species, providing a window on the RNA and promoter state using only a fluorescent reporter. In using a combination of dynamic single cell data and rigorous, model based statistical analysis, this work could fairly be said to be the state of the art for quantitative systems biology.

Summary

The literature described above shows how the dynamic control afforded by microfluidic devices, combined with traditional genetic techniques and rigorous statistical analysis, can provide a wealth of information about genetic networks and individual promoters. Though the technology available has limited the statistical power and scope of experiments, modern large scale devices look set to increase both the number of cells that can be imaged in an experiment and the duration of experiments possible [39, 50, 155].

In the remainder of this thesis, we will describe tools and analysis we have developed to make the microfluidic device used in our lab practical and quantitative. We will then apply this work to an example of gene regulation in *Saccharomyces cerevisiae* : the canonical GAL regulon.

Chapter 2

Microscope Characterisation and Fluorophore Selection

Microscopy is of course ubiquitous in cellular biology, and in the modern era the same could well be said for fluorescent microscopy. Using dyes, genetically encoded fluorophores and fluorescently labelled antibodies and oligos, fluorescent microscopy allows us to observe diverse properties and behaviour of proteins, mRNA and DNA within the cell[5, 134, 169]. To investigate transcriptional regulation we require high quality quantitative measurements of a reporter that can be used to infer transcription rate. Further, we need to make an accurate, quantitative measurement of the error in our observations in order to assess the statistical strength of any conclusions that we draw. In this chapter we will begin by discussing efforts to ensure the fluorescence measurements on our microscope are quantitative and to assess and minimise various sources of error. We will then describe the selection, construction and characterisation of fluorescent reporters well suited to investigating transcriptional regulation.

2.1 Microscope Characterisation for Quantitative Microscopy

The steps required for maximising signal and removing systematic errors in microscopy are well established [160, 187, 188]. Sample preparation and microscope configuration can both affect the signal, while certain post processing steps are required to remove systematic errors. We begin by discussing the systematic errors before addressing random errors or noise. The intensity $\bar{I}$ measured at each pixel x, y of a microscopes image is generally modelled as [187]:

$$\bar{I}(x,y) = S(x,y) \times f(x,y) + B(x,y) \tag{2.1}$$

where:

S is the shading or flat field, and is generally spatially varying due to spatial variation in the illumination of the sample.

f is the signal of interest (denoted f because we are usually interested in a fluorescent signal). Essentially what would be obtained with a perfect microscope.

B is a spatially varying background that can come from the electronics, stray light or most commonly from the media in which the sample is mounted.

It should be noted that in this analysis the brightness of the sample, f, would include the autofluorescence of any cells in the image: we are at this stage only considering measuring light emitted by the sample and not in distinguishing its biological source. For the final measurement to be proportional to the brightness of the sample (f) these systematic errors must be corrected, and failing to do

so can lead to a significant distortion of the measurement [38]. Equation 2.1 describes an idealised case, and the actual measurement made will be a sample from a distribution around this idealised $\bar{I}$. Generally the measured intensity, I, is well modelled by [121]:

$$I(x, y) \sim \text{poi}(p\bar{I}(x, y)) + e \tag{2.2}$$

Here, p is the combined probability that a photon is produced by a fluorophore in the sample and subsequently detected, which depends on both the microscope and the camera, and $\text{poi}(x)$ denotes a Poisson distribution with parameters x. This component of the noise is generally called shot noise. Obviously the Poisson distribution is an approximation, given that our signal has an upper bound, but it is generally an accurate one for the working regimes of the microscope[187]. e is generally called readout noise and is a contribution coming from the readout electronics and the camera. Though it is independent of the sample it may depend on camera settings and is generally modelled as being Gaussian distributed. Both these noise sources make it clear why brighter samples provide a better signal to noise ratio. If f is large this readout noise will obviously contribute a smaller proportion of the signal, while the standard deviation of the shot noise will also be reduced relative to the mean. This will produce a more accurate estimate of $\bar{I}$, and therefore f, for higher signals. The form of the distribution also makes it clear why a high background signal, B, is problematic even if it can be accurately estimated. A high background signal will increase the standard deviation in the measurement I, which will lead to a less accurate estimate of f even if the correct background signal, B, is subtracted.

The result of these corrections is a fluorescent image, or possibly a stack of fluo-

rescent images at different focal planes, of the cell. It is still necessary to relate these images to the physical property of interest, in our case total cellular protein content. Generally the mean or median of the pixels comprising the cell is taken as a measure of protein concentration [32, 36]. Whether this is accurate depends on the optics of the microscope, the size and configuration of the cells and the distribution of the protein [65].

In the following two sections we will present work addressing the issues discussed above for our system. We will begin by describing a novel method for measuring the flat field correction ($S(x, y)$ in equation 2.1) and direct measurements of the shot noise and detector noise for our system. We will then discuss methods to assess the fidelity of statistical operations as measures of cellular protein content, and how they can inform experimental design. All characterisation was done using our primary microscope, a Nikon Eclipse Ti inverted microscope controlled using custom Matlab scripts (Mathworks). An incubation chamber (Okolabs) was used to maintain the microscope and microfluidic devices at a constant temperature of 30 °C. Experiments used a 60X 1.2NA water immersion objective (Nikon). Images were acquired using an Evolve EMCCD camera (Photometrics) with a 512 x 512 sensor.

2.1.1 Flat Field Correction and Background Subtraction

Effective background subtraction and flat field correction can be difficult . Background can change over the course of a time lapse experiment and effective subtraction can require either the repeated acquisition of empty fields of view or careful analysis across numerous positions [27]. If the microscope is properly shielded from light sources external to the experiment, then the largest source of background is usually the media in which the cells are suspended[186]. In certain applications, like glass titre wells or conventional slides, there can be sev-

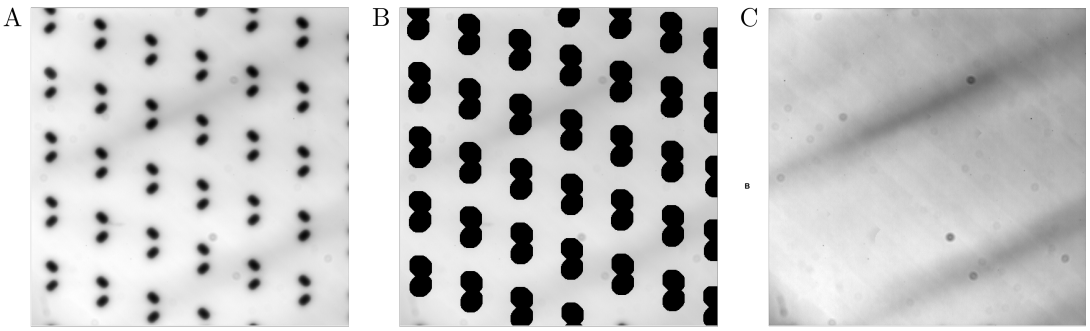

Figure 2.1: The modified flat field measurement protocol using the ALCATRAS device. A large stack of images are taken of fluorescent dye in the ALCATRAS device (A). Trap features are automatically detected and blotted out (B) and the average of each pixel across the image stack used to calculate the flat field correction (C).

eral millimetres of media above the sample which can contribute significantly to background fluorescence. Fortunately, our microfluidic devices confine media to an approximately 5 μm thickness, and the background contribution is therefore negligible.

Generally the flat field correction is measured either using either a fluorescent plastic slide [80] or fluorescent dye adsorbed onto a cover slip [160]. Though this first method is straightforward it gives an inaccurate measurement of the flat field. Fluorescent plastic slides are usually several millimetres thick and so a large proportion of the fluorescence measured will not be from the plane of focus. This is particularly true with epifluorescent microscopes such as ours, which has a lower z resolution than a modern confocal. Since in the second case the dye is located on the surface of the slide, the method is capable of giving an accurate measurement of the flat field correction, but in practice it is difficult. The dye is unevenly distributed, will begin to dissolve into the media as soon as it is mounted and is often very photoinstable,causing it to bleach significantly over the course of the experiment. This compounds the difficulty of focusing on the dye, which is usually only visible in the fluorescent channel.

We developed and alternative method using the ALCATRAS microfluidic device; the protocol is shown schematically in figure 2.1 and dscribed in full in appendix

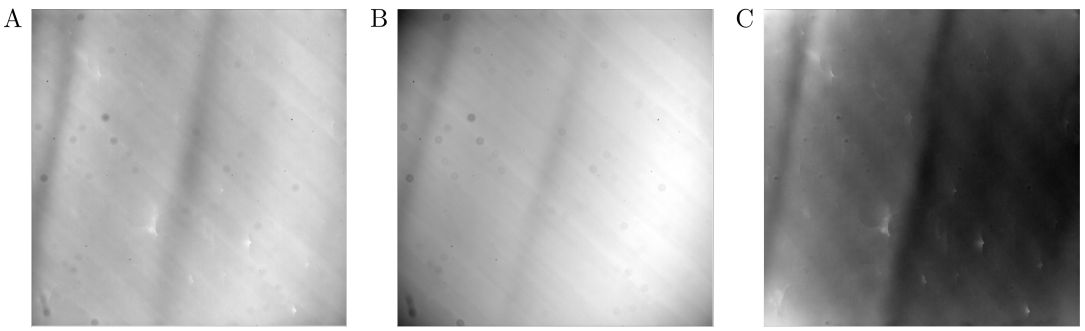

Figure 2.2: A comparison of the flat field correction found using the protocol described (A) and a y-channel device with a much thicker width (B). The difference of the two is shown in (C). Although in this case there are some artefacts from the trap removal procedure, more notable are the large scale differences, which are consistent with convolution of a much thicker section of fluorescent material. The difference between these corrections can be as much as 5% of the average - corresponding to a 5% systematic error in fluorescence.

A. No cells are loaded into the device and instead of media a high concentration fluorescent dye is flowed through, with flow set at a low rate to minimise pressure and device deformation.The device is left for around an hour to equilibrate before a frequent series of fluorescent images is taken over many position. The flat field correction is found by averaging all these images after trap features and the surrounding area have been excluded by automated scripts. To test for any affect due to the trap features or their removal the flat field was measured for several orientations of the device and a negligible difference of 0.02 % found.

This protocol has a number of advantages. Due to the trap features there is no difficulty in focusing the microscope, making the procedure straightforward. The fluorescent dye is continually replaced, so many images can be taken to get an accurate correction without the confounding effects of photobleaching or diffusion into the media. The dye is held in a narrow section corresponding to the height of the cells we wish to image, so the resulting correction is precisely that which we wish to apply to fluorescence originating within the cells and is not distorted by out of focus light. This is illustrated in figure 2.2, where the flat field measured

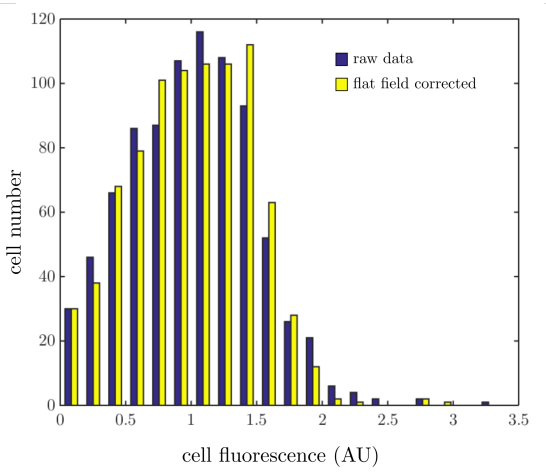

Figure 2.3: Result of application of flat field correction to cell images. A single time point was taken from a time lapse experiment in which Gal1-GFP expressing cells were grown in 2% galactose media. Cells were segmented according to the algorithms described in chapter 3 and data extracted both with and without the application of the flat field correction. The two sets of measurements were normalised by their mean and cells below a threshold size discarded as spurious cell detections. The figure shows a histogram of the two data sets, with the uncorrected data labelled 'raw' and the data extracted after the application of the flat field correction labelled as 'flat field corrected'. The flat field correction can be seen to reduce the tails of the distribution. The normalised standard deviation of raw and corrected data are 0.49 and 0.46 respectively: a drop of 5%.

by our protocol is compared with one measured by filling a much taller (50 μm as compared to 5μm) y-channel device with fluorescent dye. Although it is difficult to establish which is the 'true' flat field correction to apply, the differences are consistent with that expected from out of focus light from a thicker section, in that the flat field from the y-channel (figure 2.2.B) is a 'smeared out' version of the trap flat field correction (figure 2.2.A) with an additional hump just right of centre where the most out of focus light would be seen. The small circular features are preserved between the two; these are due to dust in the filters and as such are not dependent on the sample.

To assess the effect of the flat field correction on actual data, cellular fluorescence was extracted for a single time point of a time series with and without

the application of the flat field correction. The results are shown in figure 2.3 and described in more detail in the caption. In this case the flat field correction reduced the standard deviation of the population fluorescence distribution from 0.49 to 0.46 : a drop of approximately 5%.

The overall order of the flat field correction shows how vital it is. Even after the microscope was optimally configured difference of up to 11% were still observed in the flat field, which produced a systematic error of 5% in measurements of protein expression if not corrected. Given that the flat field correction is likely to change for different fluorescent channels, this would be particularly important to experiments comparing fluorescence in more than one channel, such as measurements of intrinsic noise or ratiometric measurements of pH [11].

2.1.2 Measurement Noise Estimation

As discussed, measurement noise is the random variability in measurements of identical samples due to the probabilistic production and detection of photons and the thermal and electronic noise in the detector and readout electronics. We use a cooled electron multiplication charge coupled device (EMCCD) camera which can operate in either CCD mode or EM mode. In CCD mode, photons incident on the CCD liberate electrons which are subsequently readout by a readout amplifier. In EM mode these electrons are first amplified by the application of a voltage gain, which causes the original electron to produce many electrons in a stochastic avalanche process, before they are readout by the readout amplifier. Since the amplification is stochastic it can increase the shot noise of the camera, but it also causes a higher signal, thereby reducing the contribution of readout noise to the final signal [145]. Given these contradictory effects on accuracy, the choice of camera mode and gain is not obvious and depends on the brightness

of the sample to be imaged. Different camera modes can also produce varied measurements for the same signal. Though the camera is designed to make the signals as comparable as possible, its signal response and noise characteristics should be checked for each camera setting used in quantitative applications [166].

To characterise and compare different camera settings we require a consistent sample with a wide range of pixel values. As before, we used a microfluidic device pumped with fluorescent dye (fluorescein in our case) to obtain such a sample: the continuous flow of fresh dye ensuring that photobleaching did not change the sample between exposures. In place of the ALCATRAS device, an older Y channel device was used. This ensured that there were no features that might affect the per pixel variance estimate. A defocused image, close to the channel wall, provided a large range of pixel values that were little affected by small changes in focus or position. These two features meant that the whole dynamic range of the camera could be efficiently investigated and behaviour under different settings compared. For each camera setting and position, fifty consecutive images were obtained and these used to estimated pixel mean and variance. These were then compiled over three to ten positions (depending on the gain) to characterise the noise response of a particular configuration of the camera. The same images were also compared between different camera settings at a given position, allowing us to assess the way a fixed sample was interpreted under different camera configurations.

The assessment of camera linearity using this procedure are shown in figure 2.4. Repeated images of a particular position were taken with different exposure times for two different camera modes, exposure time being taken as a proxy for sample brightness. Panel A establishes the linearity of the relationship between sample brightness and pixel intensity for the camera in CCD mode (see appendix A for proof) . Using this result, we test the linearity of the camera in EM mode

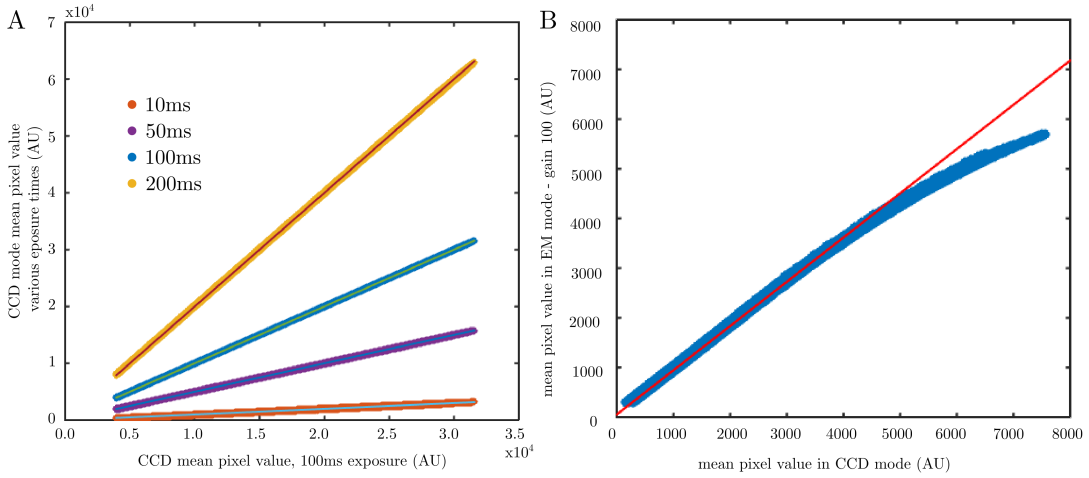

Figure 2.4: Linearity of camera response depends on camera mode. Repeated images were taken at the same position for different camera settings, providing images of identical samples with different exposure times. In panel A the linearity between sample brightness and intensity is assessed for the CCD camera mode. The mean value of each pixel is plotted for different exposure times (red is 10 ms,purple is 50 ms, blue is 100 ms, yellow is 200 ms) against the mean value of the same pixel for the 100ms exposure. As a guide, a line is also plotted for each exposure time of the expected intensity if the relationship between intensity and exposure time is linear. Panel B shows the plot of mean pixel value for the camera in CCD mode (x axis) against the mean pixel value for the same sample in EM mode at a gain of 100 (y axis). Each blue circle represents a single pixel. Given that we have established the linearity of the camera in CCD mode in panel A, a linear intensity response in EM mode would result in a linear relationship between CCD and EM measurements. Clearly this is not the case. For guidance, a linear fit to the lower range (values less than 2000 AU) of the data is plotted in red.

by imaging the same sample in EM mode and CCD mode and comparing the values. This is shown in panel B for the commonly used gain of 100, and shows that the relationship between sample brightness and measured intensity is not linear for the upper range of pixel values in EM mode. Our camera performs a built in normalisation in EM mode, so that pixel values are comparable between different camera modes and gains. This procedure is the most likely cause of the non-linearity at high intensity values. Though the implementation is unclear, EM gain generally has an exponential amplification effect [145], and an imperfect attempt to reverse this exponential multiplication could result in the non-linear effect observed.

While this is a big drawback of the EM mode of the camera, a benefit that is not shown in these figures is the absence of shading. Shading, like flat field, refers to a non-uniform behaviour of the camera across the field of view [166], but is dependent on the camera electronics and is generally only present at low intensities. In CCD mode, our camera displays shading below around 350 AU of pixel value, resulting in areas of the field of view looking entirely dark despite the presence of a fluorescent sample. This does not occur in EM mode, and is an important consideration when selecting which camera mode to use at very low sample brightness.

To quantify the dependence of pixel variance on intensity and camera settings, image stacks were obtained at a range of fluorescence intensities, exposure times and camera settings. The mean and variance of each pixel was calculated, and these means and variances were then compiled for different camera modes and gain settings. The results for both CCD and EM camera mode at two different gains are shown in figure 2.5. In CCD mode the mean variance relationship is linear, and is well fitted by the function $\mathrm{Var}\,(x) = 250 + 1.04 \times \langle x \rangle$. In contrast it can clearly be seen that the mean-variance behaviour in EM mode is not at

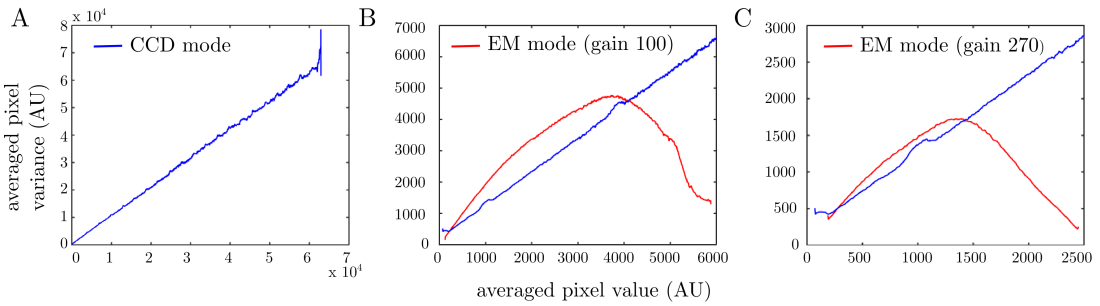

Figure 2.5: Panel A shows the mean-variance relationship averaged over three data sets for the CCD mode of the camera. Panel B shows the same, CCD mode, result in blue and the mean-variance relationship for the camera in EM mode set to a gain of 100 in red. Panel C shows the same relationship for a gain of 270 - the maximum possible for our camera. Each data set covers the full dynamic range of the camera given the settings. The averaging and smoothing was performed as discussed in appendix A.

all linear. It is strongly gain dependent and this is not due to saturation, since care was taken not to saturate the camera and no saturated pixels were found in the images. The strange, peaked behaviour is most likely a result of the built in normalisation discussed earlier, particularly since the peak occurs at a pixel intensity of around 4000 - similar to that at which non-linearity was observed in camera behaviour. This is only a hypothesis, and the physical cause of the mean-variance relationship is of secondary importance to us. Knowing the relationship allows us to select optimal settings for our experiment based on the brightness of the sample and to estimate the error in our measurement due to camera noise.

In this section we have used repeated images of a fluorescein filled y-channel device to assess the linearity and mean-variance relationship of our camera in two frequently used configurations. The results show that while in CCD mode both sample brightness-measurement and mean-variance relationships are linear, in EM mode these are both non-linear. For quantitative studies using the EM mode of the camera the brightness will have to be corrected to give accurate measurements, while any estimation of error will have to take account of the

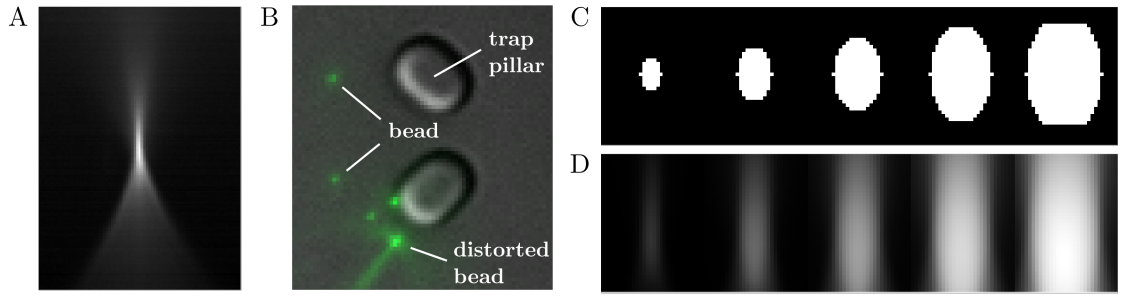

Figure 2.6: Using convolution to simulate cell observation in the microscope. The imperfect optics of the microscope distort the images acquired of samples, with light from a single point contributing to the intensity of all focal planes. This is generally represented by the point spread function (PSF): the measured intensity coming from a single sub resolution fluorescent bead. The measured PSF for our microscope, constructed as the average of the PSF measured for four individual beads, is shown in cross section in A. This PSF is distorted by the PDMS trap features, as can be seen by the overlaid fluorescent and DIC images of beads amongst the traps shown in B. In order to understand the relationship between our cells and the images we obtain of them, idealised fluorescent cells, restricted to the trap height were convolved with our measured PSF to produce simulated images of our cells in the traps. Idealised cells of various sizes are shown in vertical cross section in panel C; their convolved counter parts are shown in panel D. Though the cells appear elongated, this is an artefact of the lateral and vertical resolution used. The cells are in fact spherical, with the top and bottom curtailed if it would exceed the trap dimensions.

mean-variance relationships found here.

2.1.3 Estimating The Relationship between Measurement and Cellular Fluorescence

The above work removes systematic errors and estimates random error in the acquisition of a fluorescent image, but it still does not tell us how we should convert our fluorescent image to an estimate of the biophysical property we are interested in: total cellular fluorescence. Generally the mean or median of the pixels constituting a cell is taken as an estimate of total fluorescence concentration, and therefore protein concentration [32, 36], but there is no *a priori* reason why this should be an unbiased measure of cell fluorescence concentration. Whether it is

depends on the optics of the microscope, and in particular the point spread function (PSF): the three dimensional distribution that describes how point like light sources contribute to intensity measurements at the focal plane. This function is readily measured using sub resolution beads, and is a useful tool for detecting problems in microscope configuration [64]. The PSF measured for our microscope is shown, projected along the y-axis, in figure 2.6 A. It can be seen that there is a significant contribution from out of focus light to the plane of focus.

Having obtained a PSF, one possibility for obtaining accurate measures of total cellular fluorescence is to take dense z stacks of our cells and deconvolve them. Deconvolution is a blanket term for a number of algorithms that all aim to use the PSF and a stack of images to reconstruct the original fluorescence distribution [185] (f in equation 2.1). This procedure has a number of problems. It requires dense z stacks to be acquired, increasing the time taken to image each cell and requiring either a reduced exposure time per z slice or an increased total exposure time, which corresponds to either noisier measurements or excessive exposure to damaging fluorescent excitation light. Deconvolution is in general an ill posed problem and the fidelity of the reconstruction is not guaranteed even under ideal conditions.

To test the quantitative improvement of deconvolution under ideal conditions, we convolved our measured point spread function with a range of features and then deconvolved them using the same point spread function and the free image J plugin Deconvolution Lab [181]: an independently tested and competitive deconvolution software [55]. Even under these conditions, deconvolution could not accurately reconstruct the original features. Lastly, high fidelity deconvolution relies on having an accurate point spread function. Measurements of fluorescent beads in the our microfluidic device showed that the PDMS features distorted the PSF (figure 2.6 B), confounding any effort at deconvolution.

As an alternative to deconvolution, we sought to follow the example of Gordon et al. [65] and use convolution of idealised cells with our measured PSF to select good estimates of cellular fluorescence and to understand factors contributing to errors and bias in those estimates. Of course, the distortions in PSF due the PDMS structures will still affect the applicability of the conclusions we draw, but since we are looking more for general properties of statistical measures rather than assuming an incorrect parameter in a sensitive inversion problem, it is reasonable to hope that the effect will be less drastic. Each idealised cell was constructed as a pixelated sphere of reasonable radii and uniform value, that was capped at 2.5 μm above and below the middle to represent restriction by the ceiling of the microfluidic device. These idealised cells were then convolved with the measured PSF, giving a simulated measurement of those idealised cells at a high density of z-sections. This allowed comparison of actual total fluorescence (the number of pixels that make up the cell) and statistical measures applied to these simulated images of the cell.

In order to assess the behaviour of measures based on cell pixels, a segmentation area was defined for each cell as an accurately centred circle of with radius equal to the cells. This corresponds to assuming an accurate segmentation at the widest point of the idealised cell. An imaging plane was then chosen as the plane at which the most cells had the largest ratio of light in the segmentation area to outside it. This was chosen since, in experiments, the imaging plane is usually chosen based on maximum fluorescence contrast. In these simulations this was found to be about 0.5 μm below the centre of the cells. Four statistics were then calculated: the mean, median and sum of the pixels within the segmentation region and the sum of all the pixels in the plane. The results are shown in figure 2.7. As described in the legend, panel A is the raw value for each of the statistics and should be constant if the statistic is a good measure of cellular fluorescence concentration. Panel B is the value for each of the statistics normalised by total

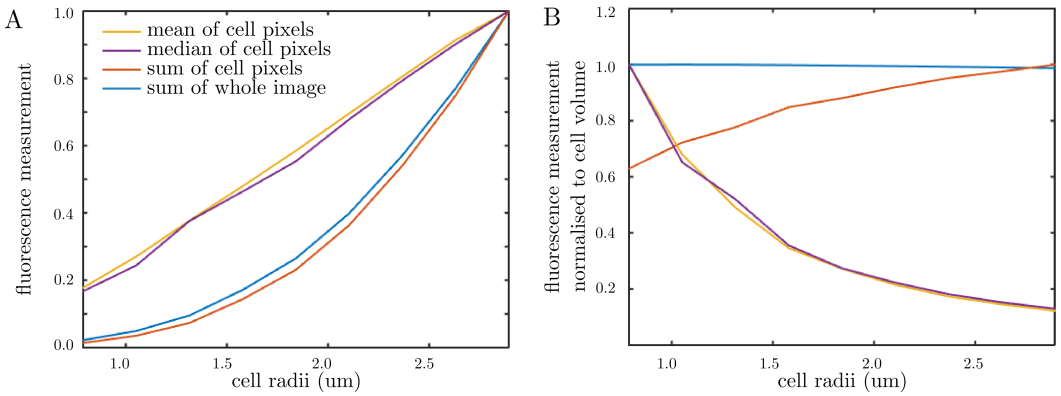

Figure 2.7: Fidelity of various measures to concentration and total cellular fluorescence. Using a single z-section from the convolved cell image stacks shown in figure 2.6 D, various statistics (mean, median and sum of cell pixels and sum across the whole image) were calculated for cells of increasing radii. Panel A shows the raw statistic normalised to their individual maxima. Since all the cells were simulated with unit fluorescence concentration, any measure which was a good estimate of cellular fluorescence concentration would have a constant value. Panel B shows the same statistics normalised to cell volume. Any statistic which is a good estimate of total cellular fluorescence would have a constant value in this plot.

cell volume, and should be constant if the statistic is a good measure of total cellular fluorescence. The range of cellular radii used is rather extreme, spanning the range from a small bud to a large senescent cell, but cells between 1.5 μm and 2.5μm are regularly seen throughout our experiments.

Since neither mean nor variance present as a straight line in panel A or B, they are both poor measures of both fluorescence concentration and total cellular fluorescence in our microscope. Total pixel sum over the whole image has a very high fidelity to total cellular fluorescence (panel B) but is of course an impractical measure, both because cells are almost always adjacent to each other in our experiments and because it would introduce a large amount of background signal from the media for dim cells. Sum of cellular pixels is a practical estimate with reasonable fidelity to total cellular fluorescence, but even when we restrict ourselves to cells in the 1.5-2.5 μm range it can have errors of up to 15 %. This is particularly important because the error dependent on cell size and would there-

fore be slowly varying and systematic. Smaller cells would appear to be dimmer throughout the experiment, leading to an overestimate of extrinsic noise and any parameters inferred there from.

The simulated images also allow us to assess the effect of focal drift, the slow change in the distance between the cells and the optics, both over time and across an experiment. To do this, the change in the ratio between the fluorescence estimate (sum of cellular pixel values) and total cell fluorescence was measured when the plane used to make the estimate was shifted by up to 1 μm : the extreme of what would be expected. For the smallest cells the effect can be significant, leading to a change of up to 13% in the ratio of measurement to fluorescence. However, for cells in the normal size range the effect is more moderate, around 3-5%, showing that cells of the same size are reasonably comparable over time and between experiments - at least as far as focus effects are concerned.

Another issue that can be addressed is the variation in measurement due to movement of cells in the z direction. The traps confine the cells to an approximately 5 μm plane, and for cells larger than 5 μm in radius this determines their position completely, but for small cells some movement is possible that could effect the fidelity of our fluorescence measurement. To assess this, the sum of fluorescent pixels was calculated assuming a correct cell outline and plane of focus, but allowing the cells to move up and down to the extent allowed by the device ceiling. The fluorescence was estimated as the sum of cell pixels for each allowed position, giving a range of values for each cell size considered. These are shown as box plots in figure 2.8 A. It can clearly be seen that for small cells the movement in z can cause large variations in the fluorescence measurement. These are difficult errors to manage, since they would be correlated in time and difficult to estimate from the images. This strongly suggest it is best to ignore cells below a threshold

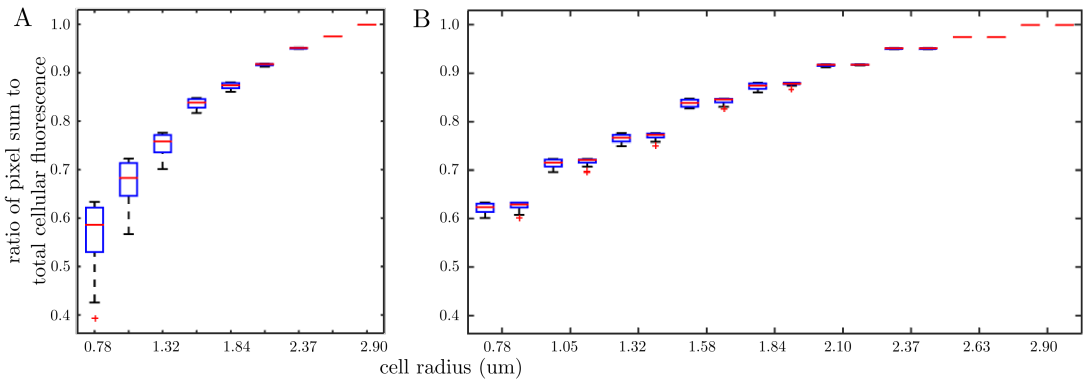

Figure 2.8: Movement of small cells in the z direction introduces a significant error. Cell images were simulated for cells uniformly distributed over the full z range allowed by their radius and the height of the traps. Assuming accurate measurement of the cell outline, the sum of cell pixels was calculated for these simulated images to give a range of values for the cell fluorescence measurement given the movement of cells. The resulting range of values obtained is shown by the box plots in panel A. For comparison, the same procedure was applied assuming that three z-sections (positioned at -1.5,0 and 1.5 μm from the central slice) and five z-sections (positioned at -1.5,-0.75,0, 075 and 1.5 μm) were taken and the maximum value of cellular fluorescence used in each case. The results are shown in panel B, with the result for three sections on the left and the result for 5 sections on the right for each radii considered. Clearly, taking z-sections produces a significant improvement in the precision of our fluorescence measurement, though 3 is probably sufficient.

radius of around 1.3 μm .

Another possible solution is to take z stacks in the fluorescent channel and use the most in focus, or brightest, image of the cell to estimate the fluorescence The effect of this procedure was simulated for 3 and 5 regularly spaced z-sections. The results are shown in figure 2.8 B. Clearly, even taking 3 z-sections and applying this procedure reduces the variation in the fluorescence measure for a cell below a certain radius. While this increased precision obviously does not change that the measurement has a cell size dependent bias, it would make any effort to correct this bias far more accurate.

The fact that total image intensity is such a good measure of total cellular fluorescence shows that the problem is not the loss of fluorescence to other focal planes, but blurring in the lateral dimensions. This results in a 'halo' around any fluorescent object, and the reason for the dependence of the bias on cell size is that this halo becomes a smaller proportion of the total fluorescence as the size of the cell increases. As shown by the fidelity of the total image intensity, over segmenting the cells (taking a larger area around the cells than the true outline) would solve the problem, but is not feasible for us since cells are often adjacent to each other. A possible experimental solution would be to use a nuclear localised fluorescent reporter, which is already commonly practised to increase signal to noise ratio [172]. By confining the fluorescent reporter to the nucleus and summing the fluorescence over the whole cell area, it may be possible to capture more of the fluorescent light emitted without capturing additional light from adjacent cells. To test this cells were again simulated, this time with all the fluorescence confined to a 'nucleus' at the centre of the cell. Nuclear radius was estimated from images of Htb2-GFP cells from the GFP collection [87] to be approximately half the cell radius for small cells and a fifth of cell radius for large cells. In our simulations the ratio of cell radius to nuclear radius was taken to scale linearly

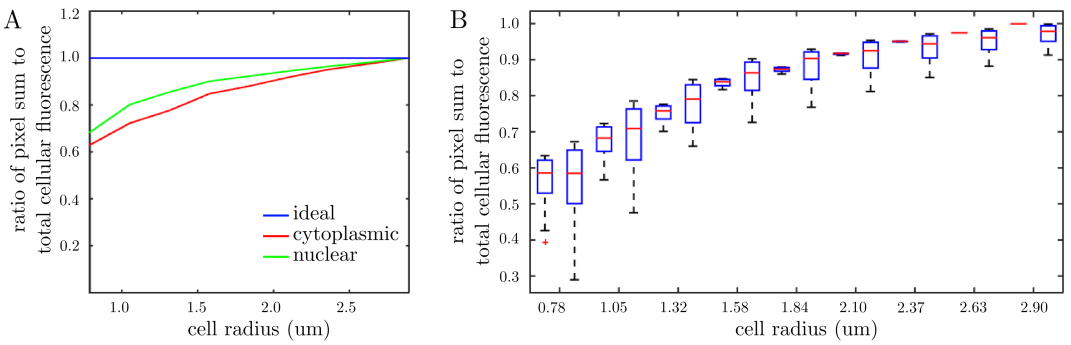

Figure 2.9: Panel A shows a comparison between the performance, on simulated cells, of the pixel sum fluorescence estimate when the fluorophore is confined to the nucleus (green line) and when it is distributed throughout the cell volume (red line). The ideal performance is shown as a guide in blue. It can be seen that the nuclear localised fluorophore gives a more accurate measurement of cellular fluorescence. Panel B shows, as in 2.8, the variation in cellular fluorescence estimate due to movement in the vertical direction. For each cell radius considered, the left box plot is the variation in fluorescence estimate when the fluorophore is distributed throughout the cell, whereas that on the right is for a nuclear localised fluorophore Clearly there is much greater variation in the case of nuclear localisation.

between these values, and the nuclei were always taken to be at the centre of the cells.

A comparison between the performance of a nuclear localised fluorophore and fluorophore distributed throughout the cell is shown in figure 2.9. It can be seen that while nuclear localisation of the fluorophore causes an increase in the accuracy of cellular fluorescence estimate (panel A), the movement of the nucleus in z can potentially cause a large decrease in the precision of the measurement (panel B). This can be reduced by taking z-section, which leads to an experimental trade off: if the signal is high, sufficient z-sections can be taken that a nuclear localised fluorophore will give a more accurate measurement of cell fluorescence without significant loss of precision; if the signal is low then the requisite high density of z-sections cannot be made without harming the cells, and the gain in accuracy will be offset by a large loss in precision.

In this section we have used measurement of optical properties with simulation to show that commonly used measures of cellular fluorescence concentration - mean and median fluorescence at the focal plane - are not robust against variation in cell size, and that pixel sum is a better measure of total cellular fluorescence. While it still has some cell size dependent bias it is more accurate, and is precise if only cells within a certain size range are considered. We have assessed the optical benefits of confining the fluorescent reporter to the nucleus and have found that there are benefits and draw backs that need to be considered in the specific experimental context.

While these errors due to the optics of the microscope are rarely discussed they clearly have a significant effect on the data. Even when confining ourselves to a fairly limited range of cell sizes we see a cell specific systematic error of up to 15 %, which will be strongly correlated in time and effect all fluorescent channels similarly. If uncorrected, this would lead to overestimation of variation between cells and would erode the validity of any quantitative conclusions based on the measurements.

2.1.4 Discussion

In this chapter we have assessed and quantified a number of distinct sources of error: uneven illumination, camera noise and errors due to microscope optics. While this is not a comprehensive list of errors that could affect our measurements, they cover those thought to be most significant in the literature [65, 186]. The importance of these errors is dependent on analysis performed on the data and the statistical significance of the conclusion, but they are not simply han-

dled. While many analyses assume that errors are independent between time points and different fluorescent channels, and similarly distributed for all cells, we have shown that some significant sources of error will be systematic errors that will vary slowly over time as the cell changes size and position in the field of view. We have further shown that common statistics applied to the cell as measurements of fluorescence concentration and total cellular fluorescence do not scale linearly with these quantities, at least not for our microscope. This has important implications for any application that requires more than a simply monotonic relationship between measurement and protein concentration.

To move further in quantifying experimental error we must first look at actual measurements, particularly of fluorescent cells in the microfluidic device, to test the conclusions drawn in section 2.1.3. Though we would expect real measurements to be more variable than the idealised ones assessed here, validating the ratiometric relationships or at least their broad trends would give confidence to an analysis that is so far largely theoretical.

That over segmented cells (i.e. the fluorescence of the whole image used earlier) provide such an accurate measure of cellular fluorescence, at least theoretically, is an unexpected and possibly fortuitous result . Though the fluorescence of the over segmented region is not a generally useful statistic, it can be applied to very bright and isolated cells. This allows us to overcome the fact that we have no 'true' value for cell fluorescence to which to compare our estimates. If the measurements of the ratiometric relationships of statistics from real cells are found to be in reasonable agreement with those predicted, we can consider taking our measurement from over segmented cells as a 'true' fluorescence for testing and refinement of measurement statistics,corrections and for estimating bias and error in our final corrected measurements.

Applying similar analysis to very bright, isolated cells with nuclear localised flu-

orophores we can make an informed decision on whether a nuclear localised fluorescent protein would give a better measure of cellular fluorescence. We already have large collections of fluorescent cells, with both cytoplasmic and nuclear localised fluorophores, and so it would only be a matter of analysing these existing images for this particular purpose.

Further measurements are also necessary of the sample brightness-measurement and mean-variance relationships, before a corrections can be found for the different camera settings used in the lab and the shot noise of our measurements estimated. Though it would require some investment to implement, our experimental work flow would then provide quantitative fluorescence measurements with estimates of a significant component of the experimental noise, a valuable result for any quantitative analysis.

It must be admitted that error characterisation is not generally thought of as exciting research, but if we want to truly understand what conclusions we can draw from our data and the confidence we can have in them, then it is vital to understand the scale and nature of the different sources of error affecting our measurements.

2.2 Fluorophore Selection and Characterisation

First isolated from the jelly fish *Aequorea victoria* in 1962, green fluorescent protein (GFP) has been a staple tool in microbiology since the mid 90's[77, 178]. In the intervening years the purification of new fluorescent proteins and mutagenesis of existing ones has expanded the range of distinguishable fluorophores [107] and improved their properties [40, 136, 153]. The requirements of a fluorescent reporter depend on the application. For inferring the dynamics of transcription it is desirable that the protein respond quickly - requiring both maturation and decay

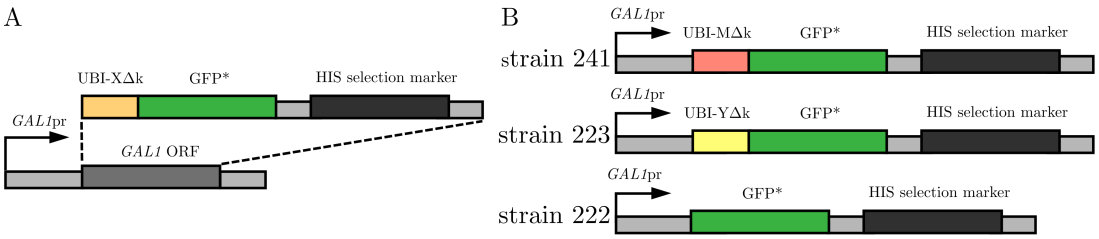

Figure 2.10: Construction of Gal1 promoter driven strains for measurement protein properties. In order to measure the properties of the transcriptional reporters engineered by Houser et al., strains were constructed in which expression of the reporter was driven by the endogenous *GAL1*pr promoter. Panel A shows the general construction of the strains. The fluorophore and selection marker were amplified by PCR, starting at the start codon of the N-Degron tag (UBI-XΔk). Flanking regions are designed such that upon transformation this sequence will replace the *GAL1* open reading frame (ORF), preserving the *GAL1* 5' untranslated region and promoter. Panel B shows the final sequence for three of the strains used: strain 241 expresses the very fast decaying UBI-MΔkGFP* fluorophore; strain 223 expresses the moderately decaying UBI-YΔkGFP* fluorophore and strain 222 expresses the GFP* fluorophore without any N-Degron tag. The N-Degron coding sequence and insertion junction was sequenced in each case and no errors found.

rates to be high - and that it is quantitatively well characterised [74, 172, 184]. Increasing degradation rate decreases the average concentration of the fluorophore, making brightness even more important to ensure a high signal to noise ratio [169]. In their 2012 paper, Houser et al. describe the development of a green fluorescent reporter designed for the inference of transcriptional dynamics. Taking as a basis the protein GFP* - a quickly maturing variant of enhanced green fluorescent protein (EGFP) [77] - they added an N terminal degradation (N-Degron) tag that marks the protein for degradation. This is effected by the proteolytic removal of a ubiquitin sequence that exposes an amino acid residue at the N terminal. The amino acid so exposed determines the rate of degradation [4, 49]. By selection of this amino acid, Houser et al. were able to engineer two fluorescent proteins: UBI-YΔkGFP* and UBI-MΔkGFP*, both having maturation times of around 14 minutes but with half lives of 120 and 10 minutes respectively.

Given that these proteins seemed ideal for our purposes, we requested two of the

plasmids described in Houser et al. [85] from Addgene - the online plasmid repository [78]. The plasmids, PNC1124 and PNC1125, contain the protein coding sequences downstream of a synthetic Gal1/Cyc5 promoter, the activity of which is strongly induced in the presence of galactose and repressed in the presence of glucose. In the original paper, the expression control so afforded was used to test the expression, brightness and decay rate of the two fluorophores. In order to repeat these experiments this same promoter-reporter sequence was amplified by PCR and transformed into the TRP1 locus of BY4741 *Saccharomyces cerevisiae* cells using a standard lithium acetate transformation procotol. Fluorescent protein expression was assayed by microscopy.

The strains expressing the fast and slow decaying fluorophores in this way were labelled strains 188 and 189 respectively[1]. The cells were impractically dim, with strain 189 (expressing the slow decaying fluorophore) displaying a fluorescence of only twice autofluorescence at full induction and strain 188 being almost indistinguishable from wild type (WT) cells. To provide a clearer signal for testing the fluorophore properties, promoter-reporter sequences for both proteins were inserted into high and low copy number plasmids [156] encoding a LEU2 selection marker. This was done using the clontec In-Fusion® system [33]. After verifying the coding and promoter sequences of these plasmids by sanger sequencing, the plasmids were transformed into BY4741 using the lithium acetate protocol.

Though the strains harbouring high copy plasmids showed a much stronger expression than strains 188 and 189 (approximately 10× the expression of strain 189), their growth was significantly slower than wild type cells when grown on galactose and fluorescence was still lower than that observed for a Gal1p-GFP fusion (approximately one sixth as bright). Given that these proteins have been engineered to degrade quickly their fluorescence is expected to be lower. However, the decay rate of fluorophore UBI-YΔkGFP* should be comparable to that

[1]Strain numbers refer to the designation in the Swain lab strain database. A complete list of all strains used is provided in appendix F

of the Gal1p-GFP fusion, implying that the steady state fluorescence of the cells expressing them should be proportional to their transcription rates. The dimness of these cells therefore indicates that, in this case, expression from the high copy plasmid is still lower than expression from the native *GAL1* promoter. This hypothesis led us to construct another set of strains: strains 223(*gal1*Δ ::UBI-YΔkGFP*) and 241(*gal1*Δ::UBI-MΔkGFP*) in which UBI-YΔkGFP* and UBI-MΔkGFP* replaced the native *GAL1* protein coding sequence and were expressed from the endogenous *GAL1*pr promoter. In addition, a third strain (strain 222 - (*gal1*Δ ::GFP*)) was constructed in which GFP* was expressed from the endogenous *GAL1*pr without any degradation tag.

The sequences and construction of these strains is shown in figure 2.10. Since in each case the fluorchrome replaces the *GAL1* protein coding sequence, these strains are unable to metabolise galactose [90]. Growth and strong induction can be achieved by a combination of raffinose and galactose, while near complete repression is observed if glucose is the only carbon source available.

To measure the decay rate of the three fluorophores an experiment similar to that of Houser et al. was undertaken: highly expressing cells were transferred to repressive XY glucose (2%) media and decreasing fluorescence observed by fixation and flow cytometry. A detailed protocol is given in appendix B. The results are shown in figure 2.11. It can be seen that the slower decaying fluorophore, UBI-YΔkGFP*, has mean fluorescence and decay rate almost indistinguishable from that of the unmodified GFP* fluorophore. The faster decaying UBI-MΔkGFP* displays a significantly reduced half life of 75 minutes, though this is considerably longer than the half life reported for this protein by Houser et al. (15 minutes). The reason for this discrepancy is unclear. The UBI-XΔk coding regions of the inserted fluorophores were sequenced and found to be identical to the sequence reported by Houser et al.. While the mRNA sequences encoded in our strains and those of Houser et al. are different - our strains preserving the Gal1 5' UTR

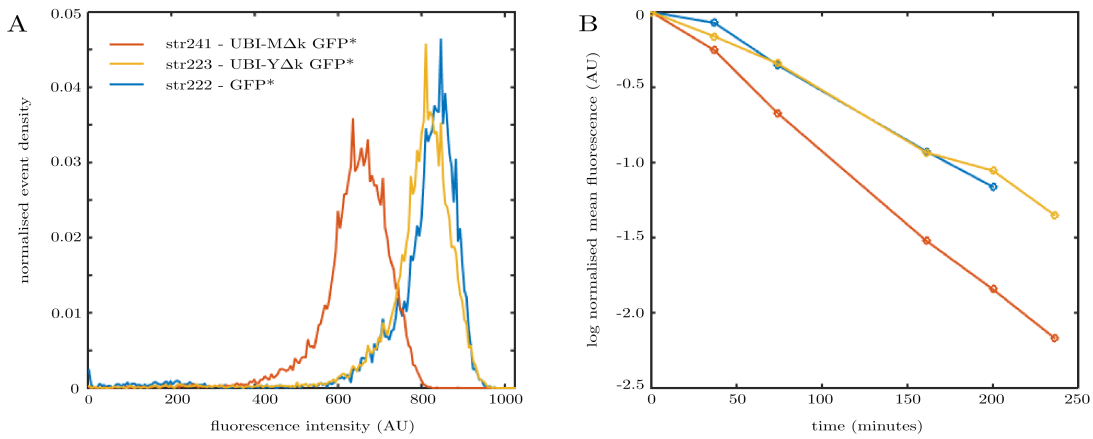

Figure 2.11: Steady state expression and decay rate were characterised for the three fluorophores from Houser et al. [85]. Cells from strain 222,223 and 241 were fully induced by growth in XY raffinose(2%)/galactose(0.1%) media, before gene expression was completely repressed by transfer to XY glucose (2%) media. Cell aliquots were extracted at approximately 45 minutes intervals and fixed for analysis by flow cytometry. Panel A shows intensity histograms for each strain at full induction. Panel B shows the logarithm of the normalised mean fluorescence for each strain plotted against time. Straight line least squares fits to these data give half life estimates of 113 minutes for GFP*, 120 minutes for UBI-YΔkGFP* and 74 minutes for UBI-MΔkGFP*. One time point has been excluded as an outlier.

while that of Houser et al. encoding a ubiquitin 5' UTR - the mRNA half life for both the Gal1 mRNA and the fluorophore mRNA was reported to be on the order of 5 minutes [85], and so would be unlikely to explain the difference. Since the proteins engineered by Houser et al. have not been widely used and no corroboration of their results have so far been published it is difficult to hypothesise on the cause of this discrepancy. However, a reporter half life of 75 minutes allows us to observe fluctuations on this timescale, which is of the order the budding yeast cell cycle and below the time scale of promoter activity fluctuations seen in both Suter et al. [172] and Harper et al. [74].

2.2.1 Constructing Fluorophores for Brightness and Fast Degradation

It is desirable to have the brightest fluorophore possible to increase the signal to noise ratio without exposing cells to excessive amount of harmful excitatory light [187]. Furthermore, it would increase the range of experiments possible if we could engineer two bright and dynamically responsive fluorescent reporters with distinguishable fluorescent spectra [74, 173]. These motivations lead us to construct two new fluorescent reporters, engineered to be fluorescently distinguishable and to mature and degrade quickly.

Having established that the UBI-MΔk motif does reduce the half life of GFP*, we sought to apply it to a pair of bright, modern flurochromes with distinguishable spectra. In their 2013 paper, Lee et al. constructed yeast optimised versions of large number of green and red proteins and assessed them for brightness, photostability and impact on cellular physiology. In experiments aimed at inferring transcriptional activity, photostability is of limited importance, since photobleaching will only contribute to effective degradation rate which we are anyway looking to increase. For this reason we can prioritise brightness and the impact on cellular physiology in selecting our reporters and largely ignore photostability. When brightness was prioritised, Lee et al. found that EGFP and GFPγ were the optimal green fluorophores of those tested while mKate2 and mRuby2 were the best performing red fluorophores.

EGFP has been widely used for many years [178] and is the work horse of fluorescene microscopy [171]. It is the foundation of many of the standard fluorophores spanning the fluorescent spectrum [107] and is still amongst the brightest green fluorescent proteins available [104]. It has been shown to mature within 25 minutes in *E. coli* [37] while its derivatives, CFP and YFP, have been measured to mature within 50 minutes in yeast at 30 °C. GFPγ was developed specifically for

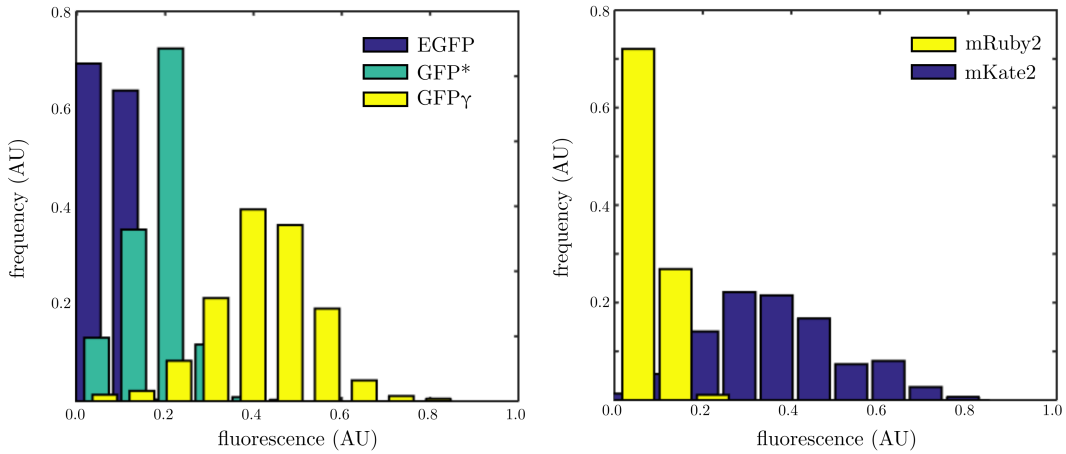

Figure 2.12: Comparison of fluorophore brightness for our system. The two best performing red and green fluorophores identified in Lee et al. [104] were obtained and transformed into the Gal1 locus of strain BY4741. Cells were grown in identical XY raffinose(2%)-galactose(0.1%) media and fluorescence assayed by microscopy. The two bar charts show histograms of the relative cellular fluorescence intensity as estimated by mean cellular pixel value. Clearly, mKate2 and GFPγ are the brightest fluorophores.

use in *Candida albicans* by combining mutations found to improve brightness in other fluorophores [197]. It was observed to be highly fluorescent, especially at 30 °C: a result later confirmed by Lee et al. [104]. Though no reported maturation rate could be found, it shares some sequence homology with superFolderGFP, which is known to mature quickly [136, 159].

mKate2 was derived by a combination of designed and random mutations of the mKate sequence [154], while mRuby2 was similarly evolved from mRuby with the intention of achieving good properties for FRET [102]. While mRuby2 is more photostable, mKate appears to be brighter [104] and mature faster [102, 154] - though these measurements, having been obtained under different conditions, are difficult to compare. Both have excitation and emission spectra well suited to the optics designed to image mCherry on our microscope.

Though the characterisation efforts described above are useful as a guide, it is difficult to know if the measurements and conclusions will apply in our case. To test which of these fluorophores would perform best in our optical system and

strains, the four fluorophores detailed were obtained from Addgene and transformed into the *GAL1* locus as described in figure 2.10. Brightness was tested by inducing all the fluorphores with a mix of raffinose (2%) and galactose (0.1%) and imaging them on our microscope. The results, shown in figure 2.12, clearly show that GFPγ and mKate2 are the brightest of the fluorophores tested. Given that there is no evidence in the literature that these should mature slower than the others fluorophores considered, they were chosen for further improvement by attachment of the UBI-MΔk motif.

The UBI-MΔk, GFPγ and mKate2 sequences were amplified by PCR and inserted into plasmids backbones with HIS and URA selection markers using the clontec In-Fusion ®system (detailed protocol in section B). The four plasmids so obtained (pFA6-UBI-MΔk-GFPγ -SpHis5, pFA6-UBI-MΔk-mKate2-SpHis5, pFA6-UBI-MΔk-GFPγ -CaURA3, pFA6-UBI-MΔk-mKate2-CaURA3) provided a convenient template for PCR amplification of both fluorophores with either HIS or URA selection markers.

As before, the two new fluorophores (UBI-MΔk-mKate2 and UBI-MΔk-GFPγ) were inserted into the Gal1 locus. Decay rate assayed by flow cytometry measurements of cellular fluorescence after the transfer from XY raffinose(2%)/ galactose(0.1 %) to XY glucose(2%) . The results are shown in figure 2.13. While UBI-MΔk-GFPγ showed a half life comparable to that of UBI-MΔk-GFP* (72 minutes and 85 minutes respectively), and significantly less than that of the unaltered GFPγ (142 minutes), UBI-MΔk-mKate2 was measured to have a half life of 164 minutes, analagous to the unaltered mKate2 (184 minutes). This is a disappointing result and limits the usefulness of UBI-MΔk-mKate2 as a transcriptional reporter. The reasons underlying the limited effect of the UBI-MΔk tag on mKate2 is unclear. It is plausable that since mKate2 is not derived from the *Aequorea victoria* GFP, its structure may be sufficiently different to interfere with the N-Degron rule. This, however, is only hypothesis and investigating the

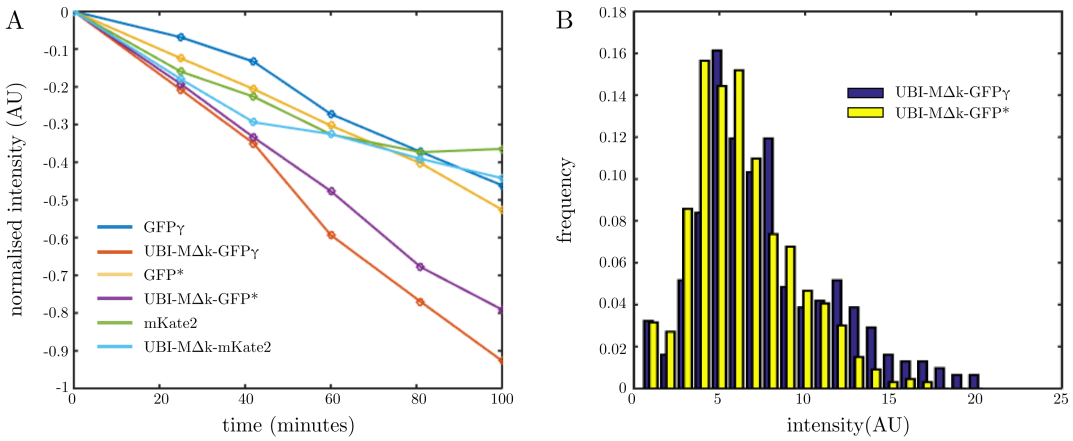

Figure 2.13: Measurement of decay rate for UBI-MΔk-GFPγ and UBI-MΔk-mKate2 with relative brightness of UBI-MΔk-GFPγ and UBI-MΔk-GFP*. As in figure 2.11, protein half lives were measured for the newly constructed proteins by flow cytometry after repression by glucose. The logarithm of the normalised mean intensity for each protein is plotted against time in panel A. UBI-MΔk-GFPγ showed a half life of 72 minutes, comparable to that of UBI-MΔk-GFP* (85 minutes) and significantly less than that of GFPγ (142 minutes). UBI-MΔk-mKate2 was measured to have a half life of 164 minutes, very close to that of unaltered mKate2 (184 minutes). Panel B shows histograms of fluorescence measurements for cells expressing UBI-MΔk-GFPγ and UBI-MΔk-GFP*. Cells were induced by growth in SC raffinose(2%)-galactose(0.1%) and measurement made using our microscope. The results show that mean fluorescence of cells expressing UBI-MΔk-GFPγ is 25% higher than that of cells expressing UBI-MΔk-GFP*.

reason for this limited effect goes beyond the scope of this PhD.
The relatively high decay rate of UBI-MΔk-GFPγ make it a useful transcriptional reporter. Measurements of it brightness shown in the right panel of figure 2.13 confirm that it is significantly brighter than the UBI-MΔk-GFP* reporter engineered by Houser et al., constituting a useful improvement.

2.2.2 UBI-MΔk-mKate2 and UBI-MΔk-GFPγ Maturation Time Measurements

For our reporters to be useful in the inference of transcription we require quantitative estimates of their maturation rates [74, 172, 184]. Since maturation is generally thought to be rate limited by oxidation [146], a common method for measuring maturation rate is to grow cells, typically *E. coli* , over expressing the fluorescent reporter of interest in an oxygen depleted environment. Cells are subsequently exposed to oxygen and the observed asymptotic increase in fluorescence modelled as a pseudo first order process. Though this method is straightforward, it requires a cellular environment that is from those used in the our experiments, and the result is only an accurate measure of the maturation rate if oxidation is significantly slower than the preceding stages of maturation. An alternative method, employed in Gordon et al. [65], is to induce expression of the fluorophore and then impede translation, by the addition of the small molecule cycloheximide, before steady state has been reached. This has the advantage that the maturation time so measured encompasses the whole period between translation and fluorescence, which is what we require to be able to infer the state underlying species. If performed on a microscope then variation in maturation rate between cells can also be observed.

In order to acquire such data for the two fluorophores under analysis we use the

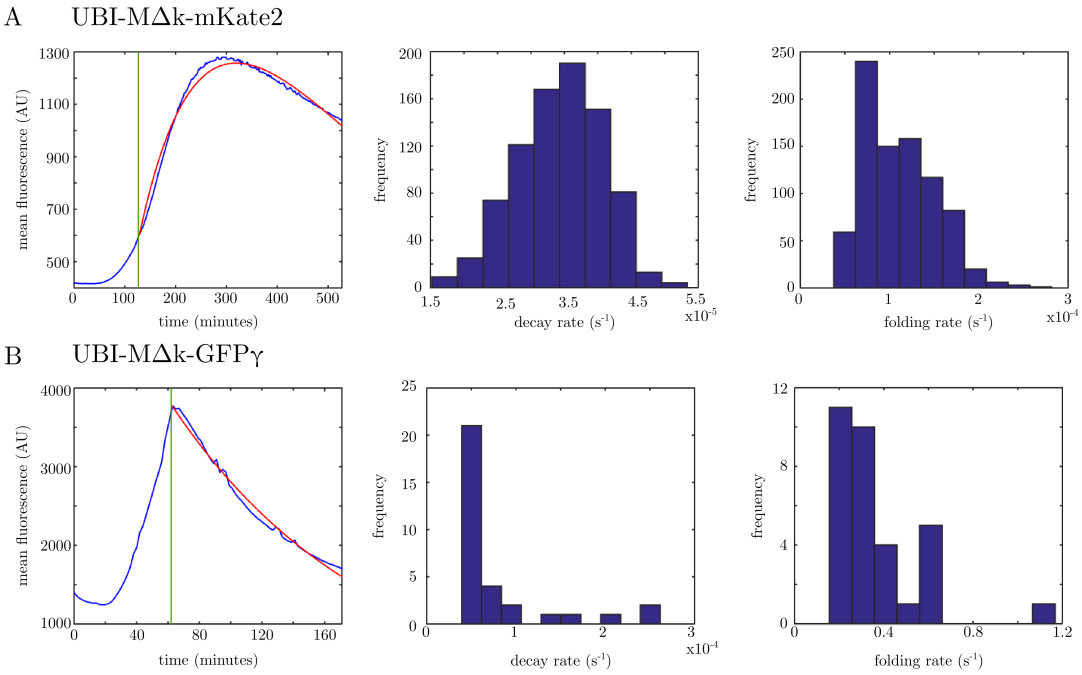

Figure 2.14: Results of cycloheximide chase experiments for UBI-MΔk-mKate2 and UBI-MΔk-GFPγ . Cells were grown overnight in synthetic complete (SC) media with 2% raffinose. Cells were rediluted in fresh media and grown for four hours before being adhered to glass bottomed Petri dishes using concavinin A. Once on the microscope the media was replaced with SC raffinose(2%)/galactose(0.1 %) to induce expression. Cells were imaged every 2 minutes and when a reasonable expression was observed the media was replaced with SC raffinose(2%)/galactose(0.1 %)/cycloheximide(20 μg / μl) to arrest translation [65]. The results are shown for UBI-MΔk-mKate2 in panel A and UBI-MΔk-GFPγ in panel B. The left most plot shows the mean fluorescence of the population in blue, with the point at which cycloheximide was added shown by the green vertical line. A simple ordinary differential equation (ODE) model of maturation was fitted to the data after addition of cycloheximide, and the results of this model are shown in red (see main text for details of the model). The same ODE model was fitted to each individual cell in order to assess cell to cell variation in the parameters. Cells were selected for a minimum fluorescence value and goodness of fit and histograms of the parameters of interest for these cells are shown in the centre and right plot for both fluorophores: central plot is a histrogram of decay rate (s^{-1}) and right plot is a histogram of folding rate (s^{-1}). These fits give a mean folding time across the selected cells of 104 minutes for UBI-MΔk-mKate2 and < 5 minutes for UBI-MΔk-GFPγ .

same strains as before. Cells grown overnight in synthetic complete (SC) media with 2% raffinose were adhered to glass bottomed Petri dishes using concavinin A, and the media replaced with SC raffinose(2%)/galactose(0.1 %). Images were acquired every 2 minutes and when a reasonable [explain further] expression was observed the media was again replaced with the inducing SC raffinose(2%)/ galactose(0.1%)/ cycloheximide(20 μg/ μl), and imaging continued. A full description is given in appendix B, and results are shown in figure 2.14. The data were fitted using the following simple model. There are two species, nascent proteins that have not yet matured, and are therefore unobservable, (x), and mature fluorescent proteins, (y). Nascent proteins mature under a pseudo first order process with rate k_f, while both mature and nascent protein degrade according to a first order process. To better constrain the model both mature and nascent proteins are assumed to decay at the same rate, k_d. These reactions lead to the following ODEs for the x and y:

$$\dot{x} = -(k_f + k_d)x$$
$$\dot{y} = k_f x - k_d y$$

These can be straightforwardly solved to give:

$$\begin{aligned} x(t) &= x_0 \exp^{-(k_f+k_d)t} \\ y(t) &= \exp^{-k_d t}\left[y_0 + x_0\left(1 - \exp^{k_f t}\right)\right] \end{aligned} \tag{2.3}$$

The equation for y was fitted to both the mean of the population and to each cell individually by minimising the square distance between the data and the model using a particle swarm optimiser. Histograms of the fitted decay and maturation rates for each cell are shown in figure 2.14. Though the fit provides a maturation and decay rate, the maturation rate is the only parameter we consider. The changes in cellular milieu due to the cycloheximide make the decay rate poten-

tially different from that in healthy cells, and we anyway have a good measure of degradation rate from the glucose measurements made earlier.

The slow response of UBI-MΔk-mKate2 to cycloheximide challenge is rather disappointing. The proteins matures very slowly, as indicated by the long period between addition of cycloheximide and peak fluorescence. The mean over the population of the fitted folding rate is $1.1 \times 10^{-4} s^{-1}$, corresponding to an average maturation time of 104 minutes. The standard deviation in k_f across cells is $3.96 \times 10^{-}5 s^{-1}$, so the folding rate is at least consistent. Though mKate2 remains a bright and useful protein tag its slow maturation rate coupled with the slow decay rate outlined previously limits its usefulness as a transcriptional reporter.

The performance of UBI-MΔk-GFPγ is far better. The fluorescence appears to peak almost instantaneously, with the maximum barely distinguishable from the first recording after cycloheximide addition. Consequently the fitted folding rate k_f is extremely fast, with a mean folding rate of 0.35 s^{-1} corresponding to a maturation time of approximately 2 seconds. This is certainly an underestimate, probably stemming from the time delay between cycloheximide addition and resumption of imaging. From the imaging times recorded by the microscope this time delay could be as much as 5 minutes, and if the maturation time were less than this most of the pool of immature fluorophores would have matured, making the folding time difficult to evaluate. This is reflected in the fitted size of the unfolded pool, x_0, which has a mean of 80 AU compared to an average starting pool of 3000 AU for folded proteins, y_0. It is clear that this discrepancy makes it difficult to accurately infer the folding rates, and fits of the mean fluorescence using the same parameters but folding times ranging from 100 s^{-1} to 1×10^{-8} s^{-1} were found to have a difference of order 1% in the least squares difference between the model and data fitted using these different folding rates. Though the data presented cannot determine the folding rate for UBI-MΔk-GFPγ, the upper bound of 5 minutes is faster than the folding rate reporter for UBI-MΔk-GFP*

[85].

Further experiments could improve our estimate of this parameter, both by more frequent image acquisition and by using our microfluidic device to switch without interrupting image acquisition. The microfluidic device was originally avoided to simplify the protocol and overcome the potential safety issues arising form its use in association with the hazardous chemical cycloheximide, but there are not technical reasons why its utilisation in combination with cycloheximide would not be possible.

2.3 Demonstration of Engineered Fluorophores in Time Lapse Experiments

To demonstrate the benefit of the fast reporter, a time lapse experiment was undertaken observing both *gal1*Δ ::UBI-MΔk-mKate2 and *gal1*Δ ::UBI-MΔk-GFPγ cells in our microfluidic device. After ten hours of induction with SC raffinose(2%) galactose(0.1%) cells were repressed by switch to SC glucose(2%). The results are shown in figure 2.15, with the shaded blue area representing the period of induction. Clearly UBI-MΔk-GFPγ is expressed earlier and decays more quickly than UBI-MΔk-mKate2, but it also appears to have more 'features': the traces appear to deviate more from what would be expected from a simple stepwise constant rate of induction and decay. Though this is difficult to quantify and is based on a visual inspection of only a sub set of the cells observed, it is the benefit we are most keen to obtain from our fast reporters: a window into underlying transcriptional dynamics that goes beyond a long period time averaged transcription rate.

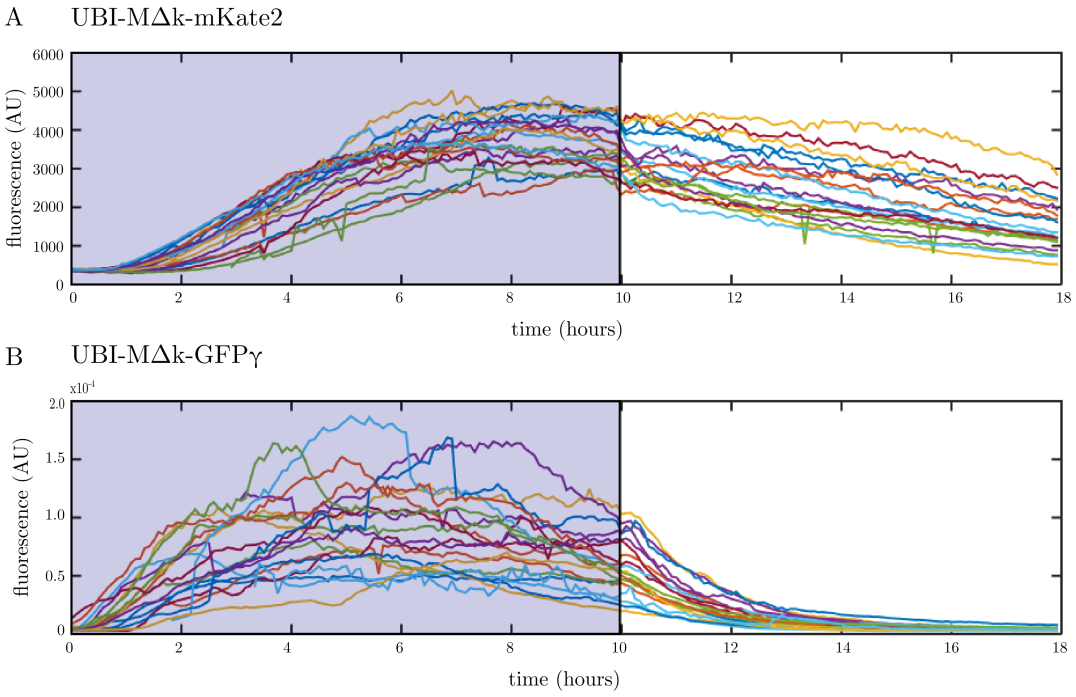

Figure 2.15: Induction and repression of UBI-MΔk-mKate2 UBI-MΔk-GFPγ in the microfluidic device. Both cells expressing UBI-MΔk-mKate2 and UBI-MΔk-GFPγ driven by the native *GAL1*pr promoter were loaded into the microfluidic device and expression induced using SC raffinose(2%) galactose(0.1%) for ten hours before expression was repressed by switching cells to SC glucose(2%). Panel A shows twenty traces for individual cells expressing UBI-MΔk-mKate2, while panel b shows twenty traces for cells expressing UBI-MΔk-GFPγ . In both case the blue area indicates the period of induction.

2.3.1 Discussion

In this section I have described the construction and characterisation of two new transcriptional reporters: UBI-MΔk-mKate2 and UBI-MΔk-GFPγ . While UBI-MΔk-mKate2 proved disappointing, being slow to fold and decay, UBI-MΔk-GFPγ displays similar kinetic properties to the UBI-MΔk-GFP* reporter engineered by Houser et al., but with improved brightness and potentially faster maturation. Use of this reporter will increase the accuracy of our fluorescence measurements, especially in cases where the protein is only lowly expressed and camera noise is dominant.

The low maturation and decay rates of UBI-MΔk-mKate2 limit its usefulness for inferring transcriptional dynamics, and confine us to experiments in which only one good transcriptional reporter is required. It is possible that other red proteins, such as mRuby2, could be more responsive to the N-Degron system employed, but constructing and testing fluorophores is a relatively long process and some evidence as to which red fluorophore is most likely to affected by the N-Degron tag would be preferable before beginning construction. The N-Degron system has been shown to reduce the half life of CFP [69], which is derived from EGFP, and so perhaps focusing on EGFP derived proteins gives a higher chance of success. CFP itself could be used, but its emission and excitation spectra have significant overlap with those of GFP, complicating their use together.

Progress could be accelerated by making use of new, high throughput facilities, recently made available in Edinburgh. The rapid, parallel DNA construction techniques of the Edinburgh Genome Foundry could be applied to add the UBI-MΔk tag to many members of the large collection yeast optimised fluorescent reporter encoding plasmids available from K. Thorn [104]. These are all provided via Addgene on identically structured plasmids, substantially facilitating the work. The resulting reporters could then be tested for fast degradation, and

to some extent maturation, using the high throughput flow cytometry equipment recently obtained by the centre. This would significantly increase the chance of finding pairs of reporters with the desired properties.

Equipped with the fast responding, bright and well characterised transcriptional reporter we have developed it, is now possible for us to undertake experiments to infer the dynamics of transcription such as those performed in Suter et al. [172] and Harper et al. [74], while our microfluidic system and automated analysis software allows us to collect data for many more cells. Our preliminary efforts to perform such experiments will be discussed in chapter 5.

Chapter 3

Automated Image Segmentation

The microfluidic device developed by M. Crane and I. Clark in the lab is essentially an array of over a thousand cell traps positioned in a flow chamber. An upstream y-channel configuration allows media in the flow chamber to be switched with high speed and accuracy [39]. A schematic of the device is shown in figure 3.1. With improvements in device loading and operation, it is now possible to observe cells in almost all one thousand of these traps for over 30 hours, and often up to 60 hours, with an imaging frequency of five or ten minutes. Obviously, such a quantity of images makes cell selection and segmentation (the process of outlining an area in the image which one deems to be a cell) by hand prohibitively time consuming. Consequently, robust, automated image processing is necessary for an efficient experimental work flow. Tracking is also required to give data over time. Given the very long experiments undertaken in the lab, even if tracking errors occur only rarely they could can accumulate to significantly erode data quality.

In this chapter I will describe algorithms developed to effectively and automatically track and segment cells. I will begin with a brief review of some the segmentation methods available, focusing on applications to microbial cells and yeast in

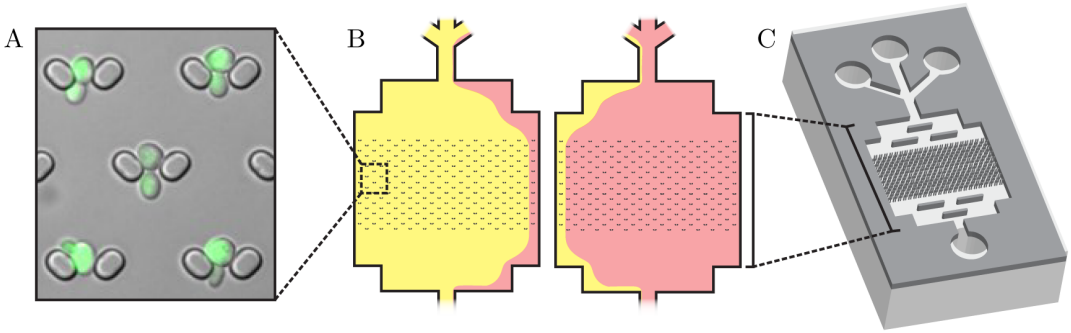

Figure 3.1: An illustration of the microfluidic device used in the lab. The device consists of an array of approximately one thousand traps (A) that can restrain a cell over multiple divisions. These are located in a flow chamber, in which the media can be rapidly switched between two sources by pressure changes (B). A 3D schematic is shown for clarity (C).(reproduced with permission from [39])

particular. I will then describe the segmentation architecture developed in the lab and present results evidencing its efficacy. The segmentation software written in the lab was a collaborative effort between multiple lab members. Though this chapter focuses on my contribution there is necessarily some discussion of the work of others, and I have tried to assign credit clearly throughout. For this reason I also break with convention and use the first person singular pronoun, 'I', at points in this chapter.

3.1 Review of Automated Segmentation Methods

I will begin by reviewing the literature on image segmentation and classification in microscopy, for which a number of reviews exist [29, 45, 116, 117].

The problem of automated cell segmentation, the process of identifying and outlining individual cells, has been a subject of research for over fifty years. The earliest segmentation methods were based on thresholding, either of brightfield images of the cells or fluorescent images generated by fluorescent dyes and stains. These

methods are still commonly used today [43, 66, 138] and are the basis of CellProfiler, a popular and widely used segmentation and image analysis platform[26]. In such methods, pixels above or below a certain threshold are deemed to be 'foreground pixels' and are therefore part of a cell. Contiguous areas of foreground pixels are grouped as cells, normally after some sort of post processing that further separates or discards areas that are too large, too small or do not conform to the statistical properties of cell like objects in some other way.

While these methods are often satisfactory they have a number of draw backs. Thresholding on brightfield images is rarely effective and as such a specified fluorescent channel has to be used. This requires cells to be stained or engineered to constitutively express a fluorescent protein, solely for the purpose of segmentation: increasing the workload for cell preparation and reducing the quality and quantity of data one can obtain. In one recent paper[32] a large proportion of the yeast ORF-GFP library [87], over four thousand cell lines, were tagged with red fluorescent protein purely so that segmentation could be automated. A further problem with threshold based segmentation is that densely packed cells are often identified as a single object. A watershed heuristic is often applied to separate them, but this prone to over segmenting [117]. While it is possible to impose more detailed shape information to prevent this [183], it is not straightforward to do so in a way that is rigorous and theoretically well founded.

An advancement on simple thresholding is to use machine learning techniques to classify pixels and threshold the results. This can be done on brightfield or Differential Interference Contrast (DIC) images, overcoming the need for fluorescent labelling. Brandes et al. [18] used local spatial and temporal variability to classify pixels as either cell pixels or background pixels. Thresholding followed by minimal morphological operations were then able to successfully segment the cells based on brightfield alone, provided they were well separated. Another example of applying machine learning methods to brightfield images comes from

Ning et al. [127]. The authors used convolutional neural networks, an advanced machine learning technique, to classify pixels in DIC movies of *C. elegans* embryos into five categories: background, cell wall, cytoplasm, nuclear membrane and nucleus. Despite the very advanced technique applied, the result still required application of a refining procedure that ensured pixels of a particular class were connected in a physically sensible way.

Active contour methods provide a straightforward and physically motivated means of encoding shape information, and have been widely used for image segmentation in many areas of biological and biomedical imaging [56, 56, 116]. In an active contour approach, the boundary is defined by a deformable contour, or snake, that is parameterised by a small number of shape parameters, Ideally, these parameters are chosen and bounded so that the contour is only able to take physically reasonable shapes. Normally a collection of seed points, generally points inside the object or objects to be segmented, are found by some heuristic and contours initialised around them. The image to be segmented is processed so that pixels likely to be part of an edge have low values and the 'best' contour is found by minimising a cost function that is dependent on both this processed forcing image and the shape of the contour. In cases where the same object in seen is numerous frames of a time lapse, the cost function can also include terms spanning numerous time points to punish unphysical changes in the object's outline over time. This can improve both the result and the computational speed by restricting significantly restricting the space of allowed contours. Whether the final contour found corresponds with what the user would define as correct depends on both the forcing image and the cost function, and in most applications the cost function will require some tuning from one instance to the next. For a more detailed description of active contour methods see McInerney et al. [116] or Blake and Isard [16].

Both Bredies and Wolinski [19] and [101] have applied active contour methods to

segmenting yeast cells. In Bredies and Wolinski [19], the contour was defined by a linear spline of thirty points and the cost function included both the derivative of the spline and a negative area term. The first was included to impose a smooth shape, while the second balanced the punishment of volume introduced by including the derivative in the cost function. The forcing image is generated based on the gradient of the image at the pixel and cell centres seeded at local maxima of this forcing image. The active contour is optimised by gradient descent starting from a small contour around the centre. Due to the inclusion of the negative area term, the gradient descent causes the contour to immediately inflate until it gets caught in the local minima created by the cell edges.

This procedure is fast and the constraints imposed on the contour mean that the result is generally reasonable. On the images for which it was designed cells are densely packed and confined to the same optical plane, so the local optimum found by moving out from the cell centre is often the desired contour. A drawback of the scheme is that forcing image takes no account of the centre of the cell, so edges of adjacent cells are equally scored. This means that in many cases the global optimum will not correspond with the local optimum found, which means the procedure relies on local optima corresponding with the users desired edge. It also makes the scheme somewhat unstable and not generalisable to cases where cells are overlapped and edges vary in appearance.

Kvarnström et al. [101] developed a similar method for segmentation of yeast, this time in slightly out of focus bright field images. Cell centres were seeded by a modified Hough transform, and the gradient along radial lines from this centre calculated. In out of focus brightfield images cells present as a white ovoid with a black halo[65], and so a forcing image was generated that was low at pixels where the radial gradient went from high to low (light to dark). The contour was optimised by dynamic programming and shape constraints enforced by transition rules in the allowed path. This procedure is more generalisable than that

of Bredies and Wolinski: the method of creating the forcing image highlights the edge in a way that is specific for the cell centre being considered and the dynamical programming method used effectively searches for global optima, which is more robust. However, the scheme only includes very limited shape information and the method for seeding cell centres often misses cells in dense clusters.

An even more advanced system for edge detection was employed in Dimopoulos et al. [42]. In this work, the authors used cross correlation of a radial trace with an edge profile, which gave extremely reliable detection of edge pixels. The authors made successive applications of a graph cut algorithm to this result to separate the image into cell and non-cell pixels. This method was chosen in part because it enforced only limited shape information, allowing the algorithm to be applied successfully to many different cell types, but this may limit the algorithms performance in poor imaging conditions where shape information can aid finding reasonable contours.

The work described above is generally developed for single images of cells and does not address the related problem of tracking cells from one time point to another. As a consequence, the segmentation algorithms described also make no use of temporal information available in time series data sets that can be used to improve segmentation. There has been a great amount of work in object tracking, and a review of tracking algorithms in cell microscopy can be found in Chatterjee et al. [29].

Tracking is usually applied after object identification and segmentation. Generally a bipartite graph is constructed by calculating a metric between cells at consecutive time points such as euclidean distance or pixel overlap. In well segmented and sparse images there are generally few conflicts and any that arise are resolved by greedy optimisation [138, 200]. Though often effective, these methods are very dependent on accurate object identification and will often fail when

objects are densely packed and misidentified. Since they act after identification, they also make no use of the history of object or biological knowledge in object identification or segmentation. A number of papers apply more advanced tracking techniques to microscopy data. Brandes et al. [18] developed an algorithm that could accurately track polymorphonuclear neutrophil cells even when they occasionally formed clusters. By uniquely identifying cells when they occurred in isolation and tracking them in and out of clusters they could ensure tracking was not lost when clusters formed. Chen et al. [30] use biological knowledge of the genesis and growth of nuclei in *C. elegans* embryos to dramatically improve the identification rate of nuclei in time series fluorescent images, and Chatterjee et al. [29] improve on the simple euclidean distance by using the previously observed motion of the object in the time lapse to track them. Massoudi et al. [114] applied an advanced algorithm from pedestrian tracking in which object identification and tracking are optimised holistically, over the whole time series, by optimising a cost function that applies an object identification potential and transfer function to all frames at once. This work shows how combining tracking and identification can lead to dramatically improved results.

Though the work described above is useful in illustrating the efficacy of methods and concepts in object identification and segmentation, it is difficult to directly apply the software produced to the images from our own experiments. Many are not designed for yeast and those that are [19, 101] must be modified to deal with the altered imaging condition, and the presence of microfluidic structures, before they can give satisfactory results. We therefore implement our own procedures using methods and ideas from the above literature. I will now discuss the algorithms so produced and present results on their relative efficacy.

3.2 Implementation of Active Contour Methods for Edge Identification

The process of image segmentation was broadly divided into three tasks: identifying cells, finding their outline and tracking the cells from one time point to the next. It was felt best to use only bright field images throughout the procedure rather than any fluorescent ones. Although more difficult, this would reduce correlated errors, reduce the work of strain construction and increase the range of experiments that could be performed[1]. M. Crane constructed a programmatic structure and implemented an algorithm by which this could be achieved. I will begin by describing this extant system before detailing my additions and presenting results. Pseudo code for all algorithms along with details of implementation are available in appendix C.

3.2.1 Algorithm 1: Centre Identification

An outline of the segmentation procedure implemented by M. Crane, and referred to as algorithm 1, is shown in figure 3.2. The traps are found in the first time points of the image by a combination of user inspection and cross correlation of the the DIC image with an example trap image. The traps are then tracked over time by cross correlation between time points, and the area around each trap used in the subsequent analysis. This both provides an absolute reference point for cell detection, that is minimally effected by drift in the field of view, and restricts the analysis to the areas of interest within the image. These DIC trap

[1]To further explain, this is because if a fluorescent image is used it must either be the same as the experimental one (thereby introducing unwanted and difficult to quantify correlations between measurements and segmentation errors) or an additional channel which would require extra transformations in any strain construction, and force us to expose to cells to more damaging excitatory light.

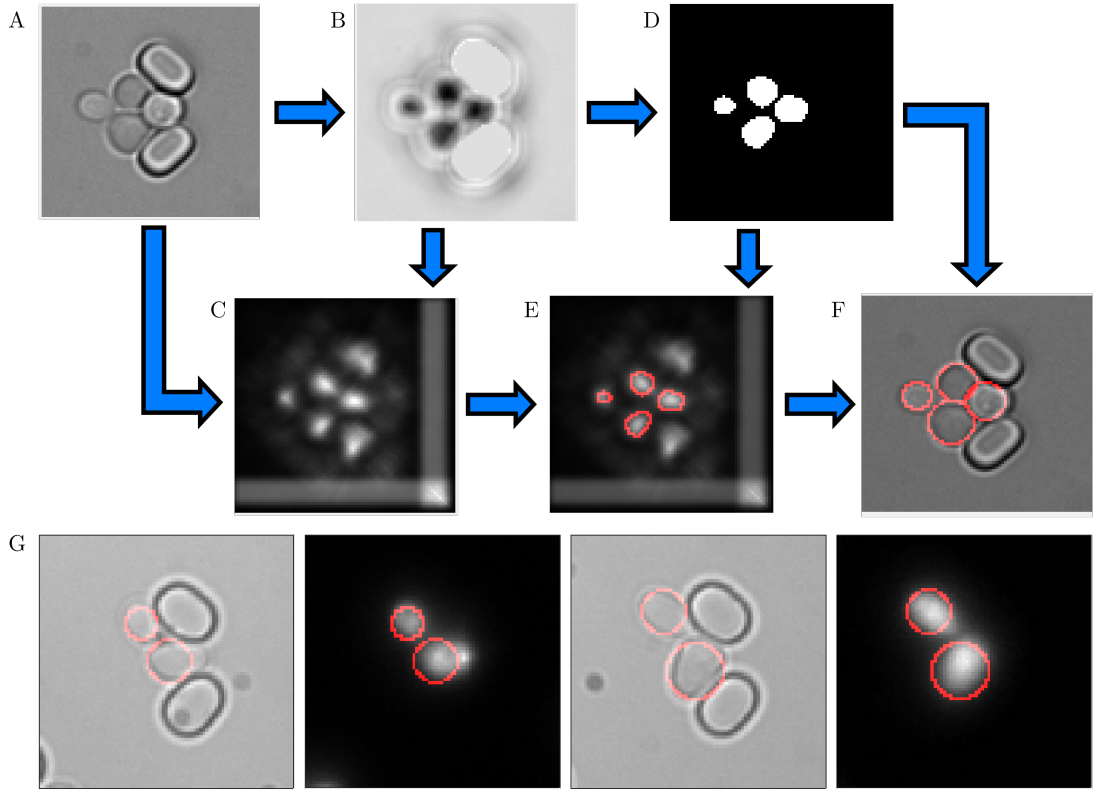

Figure 3.2: A graphical outline of the image segmentation procedure implemented by M. Crane and referred to as algorithm 1. For segmentation purposes each time point is treated individually. The image is first divided into sub-images of traps (A). A support vector machine [75] classifies pixels based on a set of 169 image transformations into a decision image (B), with low values for pixels likely to be a cell centre and high values otherwise (as can be seen, the trap is artificially removed). The decision image is thresholded into cell centre regions (D), and the average values of a circular hough transform [89] accumulation array (C) taken over each cell centre region (E). This provides a cell centre and radius which are used to construct a circular cell edge (F). Panel G shows some typical results overlaid on the DIC and fluorescence images. From these it can bee seen that though the classifier has correctly identified the cell, the circular outline misses a significant number of relevant pixels.

images are then analysed with a classifier designed by M. Crane to predict cell centres. This is the core of the segmentation software.

The classifier uses the support vector machine (SVM) [75] architecture to classify the image pixel by pixel. 169 different transformations are applied to the original DIC image, providing a vector of 169 'features' for each pixel. These are analysed by an SVM which has been trained on curated collections of cell centre pixels and non-cell centre pixels to give a cell centre score to each pixel. The image composed of all these pixel scores is referred to as the decision image.

Of the filters used the circular Hough transform[89] is of chief importance. This is a standard filter which produces high values for points at the centre of circular configurations of pixels likely to be edges(usually those with high gradient value in the image). Other features include image gradient, standard deviation between pixels in a local region and distance from a trap pixel.

The resulting decision image is thresholded using a user defined threshold and contiguous regions of cell centre pixels are determined to be a cell centre region. The final cell centre for a given cell centre region is found by taking the weighted average of the Hough transform accumulation array in this region. The edge is then found by taking the average predicted radius generated by the Hough transform for pixels in this region, and defining the edge as a circle of the predicted radius centred on the cell centre pixel. The final outlines are therefore circles, centred on points identified by the SVM classifier as cell centres.

Cells are found independently at each time in this manner and subsequently tracked. This is done by calculating a modified Euclidean distance between all cells at consecutive time points and applying a greedy minimisation of total distance. Cells are sometimes missing at single time points, which would break the tracking, and to deal with this a post processing step is applied in which the distance between tracked cells two or three time points apart is calculated, and the cells given the same cell label (identifying them to be the same cell at different

time points) if this distance is less than some user defined threshold value.
A sample of the results is shown in 3.2. It can been seen that, though the centre is within the cell and close to a user defined centre, the circular outline is often inaccurate. To improve the result, I implemented an active contour algorithm to take these identified and tracked cells and improve the outline. I will now detail this algorithm, before presenting results assessing the quantitative improvement in cell segmentation.

3.2.2 Algorithm 2: Active Contour Method for Edge Detection

From the literature described in section 3.1 it seemed that active contour methods were a good candidate for accurately identifying cell outlines. Active contour methods generally refer to methods in which a proposed edge is defined by a deformable path, which is compared to a forcing image produced in some way from the image of interest. A cost function is calculated based both on the values of the pixels along the proposed edge (usually forcing images are designed to have low values at likely edge pixels) and terms that punish unlikely shapes without reference to the image (for example, terms that punish derivatives of the path to try and force a fairly smooth shape). The combination of these cost terms is optimised for a given image, with the aim of obtaining a contour that takes account of both the image and the likely shape of the object under consideration.
I first attempted to directly apply the algorithm from Bredies and Wolinski [19], taking the centres found by algorithm 1 as seeds and making minimal modifications to adapt their procedure for our images. The result was unsatisfactory. As described in section 3.1, the algorithm used a linear spline defined by 30 points and found a local optimum for each seed by gradient descent. It seemed that for our images this local optimum was not the desired outline, and the very large

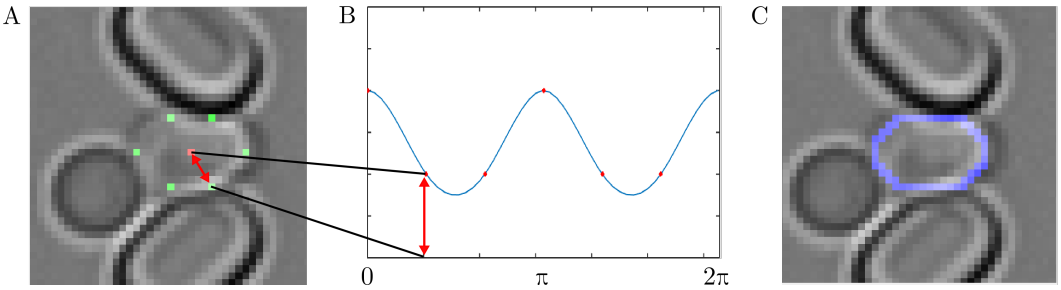

Figure 3.3: A graphical depiction of the radial spline shape space. The spline is parameterised by six points spaced at equal angles around the centre (shown in A in green with centre in red). A cubic spline with periodic boundary conditions is fitted to the radial distance of these points from the centre (B), and gives the radius of the contour at any radial angle around the centre. A dense sampling of this fitted spline is mapped back to (x, y) space and gives the final contour (shown in blue in C)

number of parameters used to define the edge (60 in a 30 point linear spline) made more sophisticated optimisation difficult. Applying ideas and methods from [16] I developed an active contour algorithm better suited to our own needs. The algorithm has the same basic components of shape space, forcing image, cost function and optimisation and I will now describe each in detail.

Shape Space

The term shape space refers broadly to the way that the path is defined by its parameters, and how these parameters are restricted to ensure that only reasonable shapes are possible. A low parameter shape space is generally desirable since it reduces the space over which the contour must be optimised, and should therefore lead to faster and more robust optimisation of the contour. Of course, too small a shape space will not allow for all cell shapes we observe. The circular outline used in algorithm 1 could be thought of as a 1 parameter shape space, the only parameter being the radius of the circle, and this appears to be too restrictive a shape space.

I used cubic splines to produce a smooth contour from a small number of param-

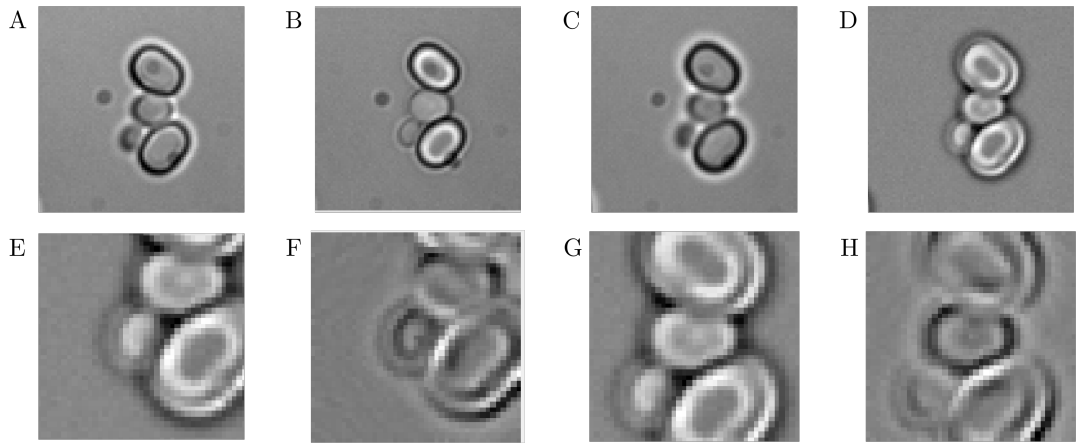

Figure 3.4: Generation of the forcing image.Shown are the DIC image (A), brightfield image 2 μm below the plane of focus (B) and 2 μm above the plane of focus (C) and the subtraction of one brightfield image from the other (D). This subtraction both highlights the edge of the cell and removes spurious objects due to dirt in the optical path. Close ups of both cells in the subtraction image are shown (E and G) as well as the application of the radial gradient image transformation described in the main text(F and H). The edge of the particular cell is preferentially highlighted, at least in the vicinity of the boundary, in both cases.

eters. I will in general refer to this method and its output as a radial spline. An outline of how the spline is constructed is shown in figure 3.3 and described in the caption. The result is a low parameter (generally 6 as shown in the figure) shape space that can fit the vast majority of cell shapes observed.

Forcing Image

In order to use the active contour structure the image of the cell has to be transformed to give low values to the cell edge. This is a little difficult with DIC images [127] since the edge has different properties depending on its orientation to the cell centre. By removing the Wallaston prism in our microscope, a more uniform brightfield image could be obtained. Out of focus brightfield images [65] proved most informative since cells appear as a clear bright shape surrounded by a dark halo below the plane of focus and as a dark object surrounded by a light halo above the plane of focus. These can be see in figure 3.4, and subtracting one

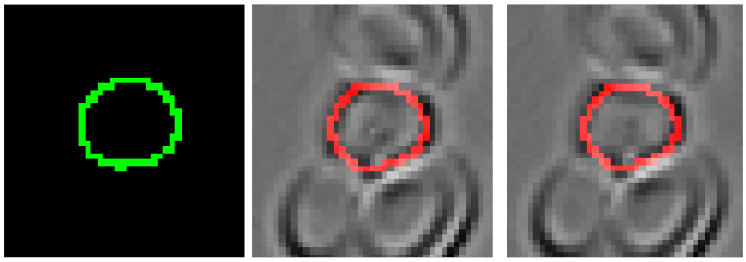

Figure 3.5: a graphical depiction of the segmentation of consecutive time points. The contour at time point $T-1$ is fixed, so the outline (green line) contributes to the cost function but is not optimised in the procedure. The forcing images for both time point T and time point $T+1$ (displayed) contribute to the cost function, and the contours for both these time points (red lines) are optimised to minimise the cost function described by equation 3.1

from the other gave an image with a clearly defined edge. Following the example of Kvarnström et al. [101] I applied a radial gradient transform to give low values to pixels at which the intensity went from high to low moving out from the cell centre. This transformation gives low values to edge pixels, but also has the advantage that it takes into account the putative centre of the cell, and favours edges belonging to that cell. This can be seen in panels F and H of figure 3.4. The pixels belonging to the trap are also receive very low scores, and these are removed by changing the value of all pixels in a predetermined trap area to be high. Details of the transformation are given in appendix C.

Cost Function and Optimisation

With a forcing image and shape space in place all that remains is to defined a cost function and the method for optimising it. Given that the cells have already been tracked in algorithm 1, it was possible to define a cost function that punished both unreasonable shapes at a given time point and unreasonable changes in shape over time. In order to do this the contour was optimised at several consecutive time points, so that the cost function can include terms dependent on the change in

outline between time points. These terms are of a fairly standard form and the final cost function, C, was defined as:

$$C(\{\vec{r}_T, ..., \vec{r}_{T+n}\}, \{I_T, ..., I_{T+n}\}) = \sum_{t=T}^{T+n} \Bigg(I_t(spline(\vec{r}_t)) + \sum_{j=1}^{m} \big(\alpha(2r_{t_i,j} - r_{t_i,j-1} - r_{t_i,j+1})^2 + \beta(r_{t,j} - r_{t-1,j})^2 \big) \Bigg) \quad (3.1)$$

where:

- T is the first time point and n time points are used for the optimisation
- $\{\vec{r}_T, ..., \vec{r}_{T+n}\}$ is the set of n vectors of length m that determine the outline of the cell at the n time points considered. In this case I used and n of 1 or 2, so either optimising the contour at a single time point or optimising two consecutive time points at once.
- $\{I_T, ..., I_{T+n}\}$ is the set of forcing images generated for the cell at time points T to $T + n$.
- α and β are parameters that weigh the various forcing terms. These are tuned for the time lapse considered.
- m is the number radial vectors used to construct the spline, in our case 6.

The first term is the forcing image term, and ensures that the outline is consistent with the images of the cell. The second is the sum of the squared differences of each radial vector with the two either side of it. This is a discrete approximate of a radial second derivative, and ensures the resultant shape is roughly circular.

The third is the sum of the squared differences between the radial vectors at one time point and the next, and punishes large changes in shape between consecutive time points.

The number of time points simultaneously considered (n) can be chosen by the user. While it may seem that increasing n would always improve the result this also increases the dimension of the parameter space over which one needs to optimise. This makes the optimisation more difficult and can result in a poorer result for the same amount of computational time. In practice, an n of one to three time points were tested, corresponding to a reasonably short time window.

Starting at $T = 1$, the cost function was optimised over time points $\{T, ..T + n\}$. When $n > 1$ was used the contour result for the last time point was discarded. This was done so that the contours assigned to the cell would always be identified considering both the time points before and after them in the time lapse. Once the contour for that block of time points had been optimised, T was set to be the next unsegmented time point, and the process repeated until the entire time lapse had been segmented. For T greater than 1, the outline of the previous time point ($\vec{r}_{T-1}$) was still included in the cost function but was kept fixed for the purposes of optimisation, so that simply contributed an extra term like the last in equation 3.1. A graphical depiction of this is shown in figure 3.5.

A number of optimisation routines were used and modification of a publicly available particle swarm optimiser [15] was found to be most effective. Although fairly 'brute force', this optimiser was good at avoiding local minima and the computational resources allocated to it could be straightforwardly increased for more difficult segmentations.

3.3 Results of Active Contour Method

Curated data sets are a gold standard for testing image analysis algorithms, and are themselves valuable for benchmarking new methods [66, 108]. In order to compare the original segmentation algorithm (algorithm 1) and the addition of the active contour routine (algorithm 2), both were applied to a set of thirty five traps from a single experiment which had been manually curated for tracking and segmentation. To curate the tracking with minimal effort images were acquired of extremely fluorescent cells, and a classifier trained on these fluorescent images. The cells were segmented using this classifier and the active contour method described in algorithm 2. Cell identification, tracking and segmentation were then manually curated using an interactive graphical user interface (GUI).
Given that we had employed the active contour algorithm being tested in constructing the base result we curate, it is possible that we may bias the results towards the active contour scheme, but no other solution presents itself. Segmenting the whole time lapse manually is simply too laborious to be practical, and were there an easier way to segment them automatically it would not be necessary to validate the active contour algorithm.

Choosing a metric on which to compare segmentation results is not obvious. There are various types of error possible such as tracking, false positive cells, false negative cells and inaccurate edge detection. The importance of these errors, and therefore the determination of the 'best' algorithm, will depend on application. To try and give an overall measure of the algorithm performance, independent of application, I applied three measures. First, the total overlap between each of the identified cells overall time between the curated and automated segmentation result was calculated and used to populate a total overlap matrix. Columns were rearranged (equivalent to relabelling cells in the automated result) to maximise diagonal elements and produce an overlap matrix that was independent of the

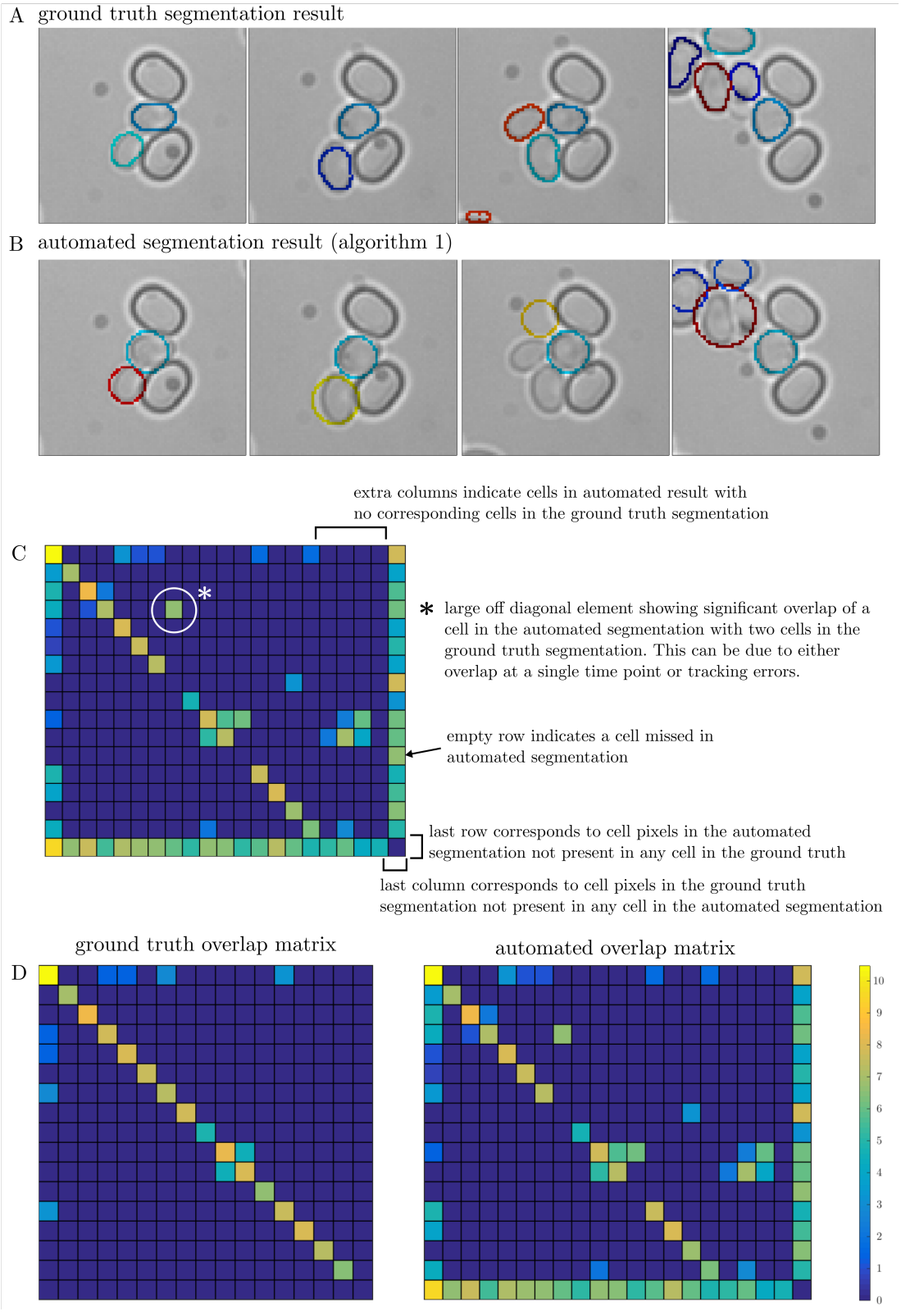

Figure 3.6: Description of the overlap matrix used in segmentation (continued on next page).

Figure 3.6: Panels A and B show corresponding time points selected from the ground truth segmentation (A) and the automated segmentation result found by algorithm 1 (B). Some common errors, like false negative, false positive and overlapping cells can be seen.The colours of the outlines are different because cell labels are not correspondent, which is not a problem for the analysis. Panel C shows the logarithm of automated overlap matrix calculated between the ground truth and overlap segmentation results for a single trap, with some description of its features. The size of the matrix shows that 16 cells were present in the ground truth segmentation, while 21 were found by the automated algorithm. Large off diagonal elements indicate segmentation errors of various types as described in the text. Panel D shows the ground truth overlap matrix and the automated overlap matrix for the same trap. The difference between these two matrices is used to calculate the summary statistics used to assess the segmentation result.

labelling of the cells. To this matrix was appended an extra row and column measuring the number of pixels in a cell that do not occur in any cell in the other result, i.e. The last column is the sum of the pixels in the curated cells that are not assigned to any cell in the automated segmentation, while the last row is the sum of pixels in the automated segmentation that do not occur in any cell in the curated segmentation.

The result, which I will refer to as the automated overlap matrix, is very informative. It identifies corresponding cells in the two time lapses without human intervention and large off diagonal elements are errors. If the overlap matrix is calculated between the ground truth and itself, a result which I will refer to as the ground truth overlap matrix, the result is a square matrix with very large values on the diagonal and small off diagonal elements where cells overlap in the image. A depiction of the ground truth overlap matrix and automated overlap matrix for a single trap is shown in figure 3.6.

As a general assessment of the result, the ground truth overlap matrix is subtracted from the top left corner of the automated overlap matrix and the sum of the absolute values of the resulting matrix, normalised by the sum of all the elements in the ground truth overlap matrix, is calculated. I refer to this as the **total error**. It can be seen that the addition or removal of any pixels relative to

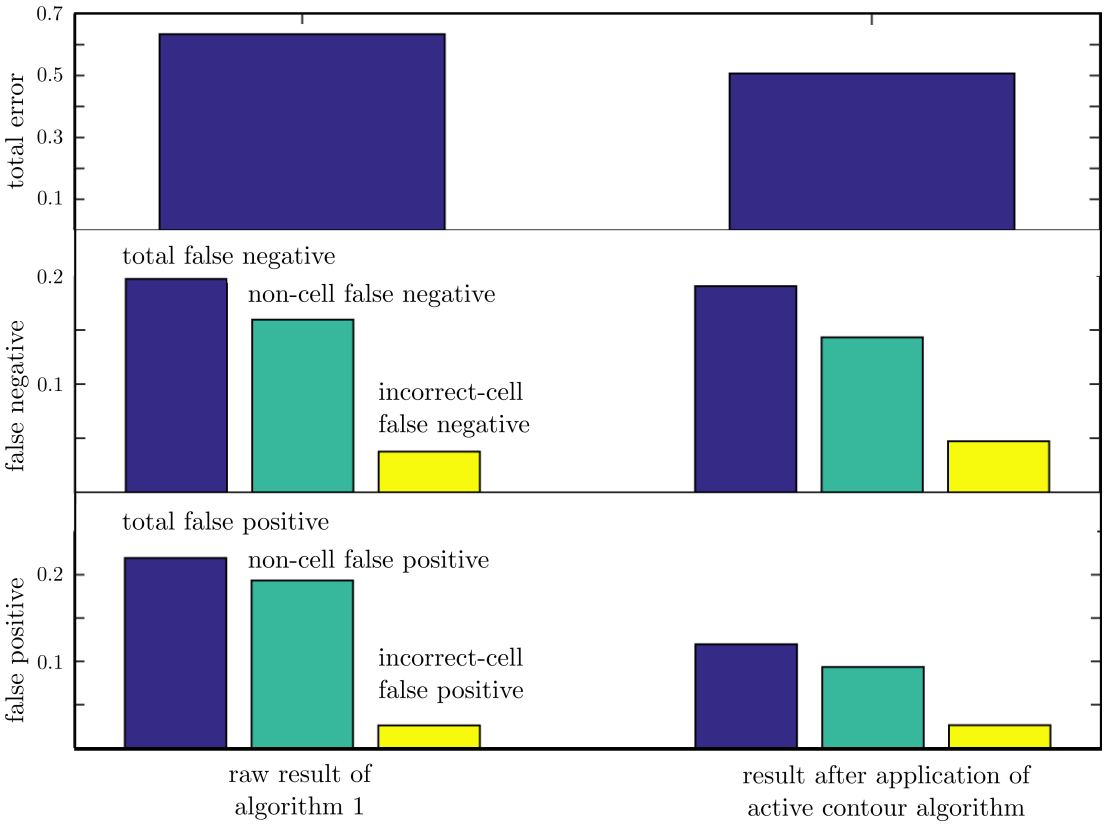

Figure 3.7: Various error measurements for both segmentation algorithms. As in the main text, algorithm 1 is M. Crane's algorithm with edges found as circles. The second result is for the application of the active contour algorithm (algorithm 2) to this result. The difference of two out of focus images were used to construct the forcing image as shown in figure 3.4. The cost function was optimised one time point at a time ($n = 1$ in equation 3.1) with a contribution from a fixed outline for the cell at the previous time point. For false negative rate we show the total false negative rate(blue), those pixels missed entirely(green) and the pixels erroneously assigned to another cell(yellow). For the false positive rate we show the total false positive rate(blue), non-cell pixels erroneously assigned to the cell(green) and the pixels erroneously assigned to another cell(yellow).

the ground truth segmentation will contribute positively to this total error. The top panel of figure 3.7 shows a bar plot of the total error both of the raw result of algorithm 1 and after the application of active contour method described in algorithm 2. The addition of the active contour algorithm reduces the total error from 0.63 to 0.5, a substantial reduction.

In addition to this total error, other informative statistics can be calculated from the automated overlap matrix. For our applications we are most interested in fluorescence measurements over the length of the time lapse, which corresponds approximately to those cells which have the highest total area over the whole time lapse in the ground truth segmentation. Focusing on the fifty cells with the largest total area (approximately ten percent of the cells considered) I calculate two further statistics. One is the fraction of the ground truth cell pixels that do not appear in the equivalent cell in the automated result normalised by the total number of pixels in the ground truth cell over the whole time lapse. The average of this fraction over the fifty most significant cells is termed the **false negative rate**. By looking at the number of these pixels that do not overlap with any cell in the automated result, we can further sub divide this into false negatives in which the cell pixels were not identified and those in which they were assigned to the wrong cell (such as in the case of tracking errors).

I similarly define a **false positive rate**. To calculate this I take the fifty columns of the overlap matrix corresponding to the cells with highest area in the ground truth result, subtract the equivalent region of the ground truth overlap, normalise each column by the total area of that cell in the automated result over the whole time lapse and sum the off diagonal terms. This corresponds to the fraction of cellular pixels in the automated time lapse that do not occur in their equivalent cell in the ground truth time lapse. This can be further subdivided into those that overlap with other cells in the ground truth, and those that overlap with non-cell pixels in the ground truth. The average of this fraction over the fifty significant

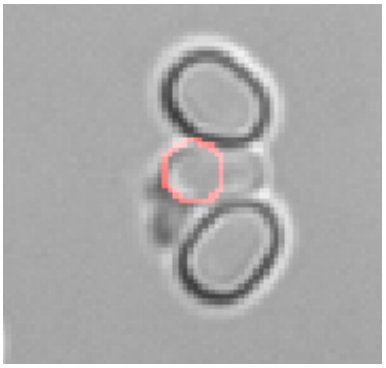

Figure 3.8: a common error occurring in the segmentation, particularly for cells in the trap. Due to the reliance on the circular Hough transform the 'centre' of elongated cells will often occur at one or other end. Even after the active contour method is applied this often results in truncated cell outlines.

cells is deemed the false positive rate. Bar charts of all these quantities are plotted in figure 3.7 for both the raw result of algorithm 1 and the segmentation after the application of the active contour algorithm described in algorithm 2.

It can be seen that for all three measures selected the active contour algorithm improves the result. The most significant improvement is in the false positive rate, and particularly the non-cell pixels wrongly assigned to be part of a cell (green bar, lowest panel). This is most probably because the circle based edge construction will tend to add extra pixels in the case of slightly elongated cells. Though the false negative rate was reduced the result was not significant. Closer inspection showed that this was most probably due to misidentification of the exact centre location. The reliance of the classifier on the circular Hough transform tends to result in elongated cells being assigned a centre at one or other end of the cell, and the resulting circular edge will only encompass around half the cell. This is not solved by the active contour algorithm since its cost function contains terms that force the outline towards a circle, so though the result is improved a large proportion of the elongated cells are wrongly outlined. An example of this is shown in figure 3.3.

This behaviour can also result in inaccurate tracking as the centre jumps along the cells length in an unphysical manner. Even if the tracking is not broken, this movement of the centre along the cells length can degrade the active contour result, which assumes the centre occurs at the same point in the cell at consecutive time points.

To try and prevent these errors and improve the segmentation result an alternative tracking and segmentation algorithm was implemented based on the methods described in Blake and Isard [16]. This was an effort to better use information about the cells at one time point in the identification of cells at the next. The algorithm is described in appendix C, but since it did not improve the segmentation result I will not discuss it here.

3.4 Error in Data Acquired Due to Segmentation

While these general errors in segmentation are of interest in assessing and comparing the various algorithms implemented in the lab, for the purposes of inference we are most interested in the errors in the final fluorescence measurements. In the data set curated, the cells are expressing a Gal1p-GFP fusion and experiencing 2% galactose in the media. As such they are extremely fluorescent, and the fluorescence can be extracted for the curated cell outlines and each of the putative cell outlines and compared. This was done, and the sum of the pixels assigned to the cell taken as a measure of cellular fluorescence (the validity of this measure in discussed in chapter 2). Since we are most interested in the cells that appear in the time lapse for long periods (generally those at the centre of the traps), we focused on the same fifty cells in the ground truth with the largest total area and their corresponding cells in the automated identification. These are the same sets of cells used in assessing the false positive and false negative rates before, and each of them is present for at least fifty time points in the ground truth time lapse.

A first observation that can be made is the number of time points at which a cell

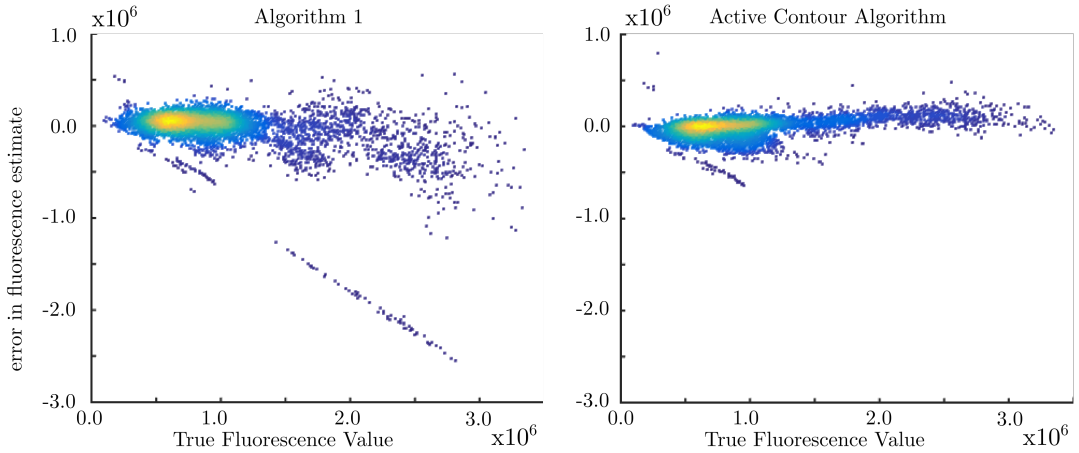

Figure 3.9: Scatter plots of the difference between the fluorescence measurement for the ground truth and automated time lapses for the fifty most significant cells in the time lapse. fluorescence measurement was calculated as the sum of the values of the pixels assigned to the cell at each time point. The scatter plots show that, particularly for high fluorescence values, the active contour algorithm significantly improves the error in the measurement.

is erroneously absent or presents compared to the ground truth. Since the two algorithms discussed differed only in the application of the active contour method to identified cell centres, cell identification and tracking results were identical in the two cases. This means that both had the same false positive and false negative rates of 4.3% and 4.7% respectively. While the false negative rate, which corresponds to lost data about, is regrettable the false positive rate is potentially disastrous, since it corresponds to entirely erroneous data. In cases of false positive assignment the data extracted will come from either another cell confused for the cell of interest, or from an empty area of the field of view. This can cause extremely large errors in the estimated fluorescence, and indicates that until the tracking and cell identification is improved it will be necessary to continue to curate at least the tracking. Figure 3.9 shows density plots of the error for each of the two segmentation algorithms. For clarity the false negative and false positive points have been removed so that only those time points for which both a ground truth and an identified cell are present feature in the plot. It can be seen that, particularly for the large fluorescence values, the application of the active contour

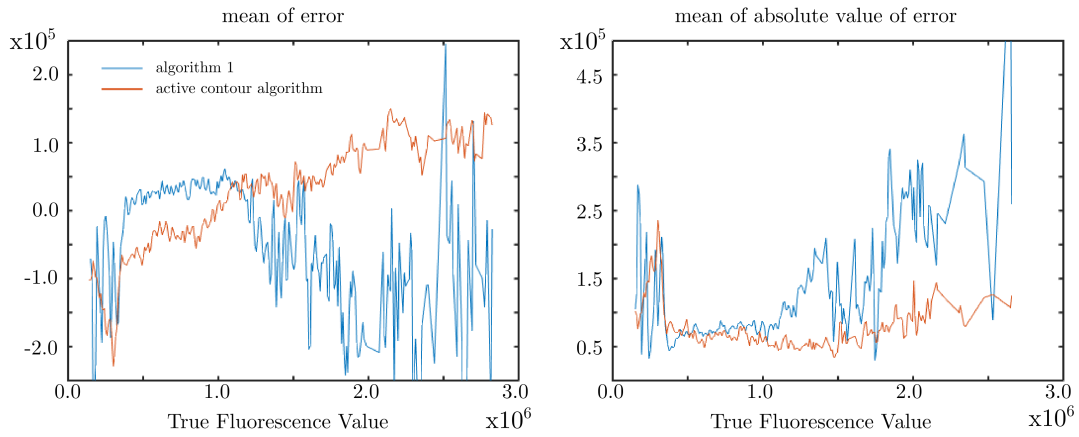

Figure 3.10: The mean and standard deviation of the segmentation errors plotted in figure 3.9. Outliers were excluded, and the mean error and the mean absolute error calculated, for a particular true fluorescence value, using all data points with true values within a small window around the desired value. This calculation was performed for data obtained using both the cell outlines found by algorithm 1 and those found by applying the active contour procedure to the result of algorithm 1.

algorithm significantly reduces the errors. It can also be seen that algorithm 1 has a larger number of significant outliers than the active contour result. To better understand the quantitative errors due to segmentation the outliers were removed and a window average used to estimate the mean of the error and the mean of the absolute value of the error as a function of the true intensity. The results are shown in figure 3.10.

There is not really sufficient data, especially at the high fluorescence values, to fully quantify the error but two trends can be clearly seen. The first is that while the mean of the absolute error for algorithm 1 scales strongly with true value, for the active contour algorithm the mean absolute error is a near constant 5×10^4. If the cell pixels were uniformly bright, and the algorithm were missing a certain number of pixels independently of cell brightness one would expect the absolute error to scale with intensity. As this is not the case, it is likely that many of the errors correspond to the addition or subtraction of edge pixels that a small contribution to the fluorescence. The second observable trend is that the mean of the error for the active contour algorithm is not zero, but shows a positive

trend with true cell fluorescence. This corresponds to brighter cells being over estimated and dimmer cells being underestimated. The reasons why this would be are unclear, and their investigation might help us understand how to improve our segmentation algorithm, but it is independently an important observation since such systematic errors can be detrimental to accurate statistical analysis.

3.5 Discussion

The segmentation and tracking of dense packed object in images is notoriously difficult. In this chapter I have presented an active contour method, a significant contribution to the existing segmentation algorithm in the lab, that attempts to tackle this problem. What is perhaps more important, I have formulated and applied an analysis pathway that allows us to assess in a consistent and quantitative way the efficacy of our segmentation algorithms, both in general terms and for our specific purpose of obtaining accurate fluorescent measurements over long time series. As a result, we are able to see that by both criteria the active contour algorithm produces a significant improvement.

Further, we have seen that segmentation errors make a substantial contribution to errors in our fluorescent measurements. False positive tracking errors are common, occurring approximately 5 % of the time, and result in such significant fluorescence errors that it will be necessary to correct them by hand. When the tracking is accurate, the modified algorithm including the active contour routine results in an error of around 5×10^4 arbitrary unit, corresponding to a modest $1 - 5\%$ error for the cells considered in this validation. This information is vital for assessing the reliability of any quantitative inference or statistical analysis performed using the time lapse data obtained from our microscope.

Going forward there are still many areas in which the segmentation software and associated analysis could be improved. The accuracy of the software is still far from that required for completely automated segmentation, and the combination of significant error rate and large data sets makes full curation extremely laborious. There are many avenues by which the software could be improved. The algorithm currently makes little use of cell information at one time point to identify cells in the next. Though efforts to apply this information have not so far resulted in any improvement, it is reasonable to hope that with refinement they will.

The shape space model currently used is crude, heuristically punishing non-circular shapes that change over time. Using curated data sets it will be possible to construct a statistically informed shape space, preventing the segmentation algorithm selecting unrealistic cell shapes when the edges are unclear or the image crowded. This modification has the advantage that it is transferable between different imaging environments, allowing us to use a shape space acquired on good images to segment poor quality images.

Beyond these relatively straightforward modifications, there are more advanced techniques that could be applied to the segmentation of our images. The support vector machines used in our classification have been the work horse of machine learning for many years, but more sophisticated classification methodologies now exist, particularly for image analysis[119, 127]. The application of these could result in substantial improvements.

Any of these avenues could improved performance, but judging any improvement by eye is difficult for any but the most substantial gains. That is why it is important that we now have a frame work in which the effect of any modifications can be rigorously assessed.

Computational cost is another area in which the software could be improved. As the microscopy and microfluidics in the lab have improved the data sets have be-

come larger and larger, with some now consisting of over a thousand traps imaged for over a thousand time points. With the current software these can take almost a week of constant computer use to analyse. Recoding our algorithms in lower level languages is always a possibility for improving performance, but is likely to be laborious and hamper innovation. A better approach may be to try and find more efficient optimisers for the active contour algorithm, and a reduced set of features for image classification. Currently only the DIC image is used in classification, and it seems likely that a more efficient classifier could be constructed using the two bright field images currently used in edge detection. Again, it is important that we now have a system in place to rigorously assess the result of such modifications.

In terms of error analysis, it would be valuable to assess performance on a larger range of data sets, encompassing more conditions and expression regimes. For example, the cells used here were very bright, but in cases where the cells are dim the contribution of false positive and negative pixels to the fluorescence measurement is likely to be different. Though curating data sets is laborious, it will anyway be necessary for those intended for quantitative analysis and publications. By preserving the result before and after curation, it will be straightforward to subject them to the same error analysis methods described here.

We could also extend our error analysis to other features of interest. Total fluorescence is not the only measure of interest in the lab and localisation of proteins, both to the nucleus and membrane, is also studied and quantified. These will likely have very different error statistics, and it would be interesting to apply the analysis described here to these measures. The results obtained could also help guide our experimental design by establishing which measures are least error prone.

Though there are many possible improvements that could be made to the segmentation software one should try to maintain perspective. Image analysis is a

difficult problem and it is unlikely we will ever have a perfectly automated system. Efforts to improve the it should be balanced against the likely benefit in human time, and one should always remember that in the end it is scientific insight that we pursue, not satisfying image segmentation.

Chapter 4

Estimation of Protein Concentration by Analysis of Stochastic Fluctuations in Photobleaching

A parameter of our experimental system that has not so far been addressed is the ratio, ν, between measured fluorescence and the actual protein content of the cell. Since the actual number of proteins present determines the size of stochastic fluctuations [135], knowing this parameter is vital for any analysis that attempts to leverage fluctuations in single cell fluorescence to draw quantitative conclusions. There are a number of ways available to estimate ν or measure it directly. Fluctuation correlation spectroscopy (FCS) [28, 113] can provide a measurement of ν but, this requires considerable expertise and specialised equipment. Another possibility is to calibrate fluorescence measurements against measurement of total protein content by quantitative western blot [195], but performing quantitative western blots is non trivial [192], and the accuracy of using publicly available

measurements [60, 111] is difficult to assess.

A number of papers have looked at using single cell fluctuations to estimate ν. Rosenfeld et al. [149] transiently expressed a fluorescent reporter in *E. coli* and observed its dilution due to growth following repression. At each replication the fluorescent protein was equally likely to locate to one or other cell, resulting in a even binomial distribution for the number of proteins inherited. The variance of this distribution was dependent on the number of proteins partitioned, and this allowed both ν and the measurement error to be estimated. Teng et al. [175] applied a similar analysis to cells displaying constant protein expression. Fluorescence was measured immediately before and after division, and variations in the size of the two daughter cells included in the analysis to give a more complete decomposition of the sources of variation.

Though both these studies successfully estimate ν , it is not clear how one could easily combine them to estimate both ν and the measurement error while allowing for the mother-daughter size disparity seen in budding yeast cells. An alternative but similar approach for estimating ν has been suggested by Nayak and Rutenberg [124]. In a theoretical paper, the authors show that single cell measurements of photobleaching - the process whereby fluorophores cease to fluoresce when continuously excited - can be used to generate stochastic fluctuations that allow ν to be estimated. Though they do not apply the methods they develop to real data or explicitly take account of measurement error, a similar analysis by Finkenstädt et al. in mammalian cells challenged with cycloheximide and actinomycin was able to give reasonable estimates of both ν and the measurement error [51].

In this chapter, I will describe our efforts to similarly estimate ν and measurement error from time series of single *Saccharomyces cerevisiae* cells undergoing photobleaching. Using Baye's law we calculate a posterior for the two parameters of interest, assuming a pseudo first order process for photobleaching, and apply it to real and simulated data. This model is found to be inappropriate for

highly expressed fluorophores, and a low expression regime is instead explored. This requires a modified analysis to take account of autofluorescence, and this is addressed in a number of ways.

4.1 Bayesian Estimate of Fluorescence Protein Ratio, ν , Without Measurement Error

We begin by investigating the case with no measurement error. We take a Bayesian approach [112] and calculate a likelihood for ν , given a set of data and a model for the underlying biophysical process. We assume the measured fluorescence intensity, I, is proportional to the protein number in the cell, n , with the proportionality constant ν :

$$I = \nu n \tag{4.1}$$

We have a series of $M + 1$ observations, $\{I_0, ... I_M\}$, generated by a series of protein concentration, $\{n_0, ... n_M\}$, which are related to each other by the simple Binomial probability.

$$p(n_{i+1}|n_i) = B(n_i, n_{i+1}, p) \tag{4.2}$$

where p is the probability that a fluorophore is not bleached between two consecutive observations.

Using the same results derived by Nayak and Rutenberg, we can relate the distribution of observation, I_i, to the known distribution of the protein, n_i, as follows:

$$
\begin{aligned}
\mathbb{E}(I_{i+1}|I_i) &= \nu\mathbb{E}(n_{i+1}|n_i) \\
\mathbb{E}(I_{i+1}|I_i) &= I_i p \\
\mathrm{Var}\,(I_{i+1}|I_i) &= \nu^2 \mathrm{Var}\,(n_{i+1}|n_i) \\
\mathrm{Var}\,(I_{i+1}|I_i) &= \nu I_i p(1-p) \qquad (4.3)
\end{aligned}
$$

Applying the law of large numbers, we obtain a normal approximation of $p(I_{i+1}|I_i)$:

$$
p(I_{i+1}|I_i) \approx \frac{1}{\sqrt{2\pi p(1-p)\nu I_i}} \exp\left[-\frac{(I_i p - I_{i+1})^2}{2I_i \nu p(1-p)}\right] \qquad (4.4)
$$

Using this expression and Baye's law we obtain for the posterior:

$$
\begin{aligned}
p(\nu, p|\{I_0, ...I_M\}) = &\left(\frac{1}{2\pi p(1-p)\nu}\right)^{\frac{M}{2}} \frac{1}{\prod_{i=0}^{M-1} I_i^{\frac{1}{2}}} \\
&\exp\left[-\frac{1}{2\nu p(1-p)} \sum_{i=0}^{M-1} \frac{(I_i p - I_{i+1})^2}{I_i}\right] \frac{p(\nu, p, I_0)}{p(\{I_i\})} \qquad (4.5)
\end{aligned}
$$

To further investigate the behaviour of the posterior, we differentiate the logarithm of the likelihood $L = p(\{I_i\}|\nu, p)$ and identify the modal combination $(\nu_{\text{modal}}, p_{\text{modal}})$:

$$
\begin{aligned}
p_{\text{modal}} &= \frac{\sum_{i=0}^{M-1} I_{i+1}}{\sum_{i=0}^{M-1} I_i} \qquad (4.6) \\
\nu_{\text{modal}} &= \frac{S}{M p_{\text{modal}}(1 - p_{\text{modal}})} \qquad (4.7)
\end{aligned}
$$

where:

$$
S = \sum_{i=0}^{M-1} \frac{(I_i p_{\text{modal}} - I_{i+1})^2}{I_i}
$$

(details are given in appendix D). This is the same estimator for ν found by

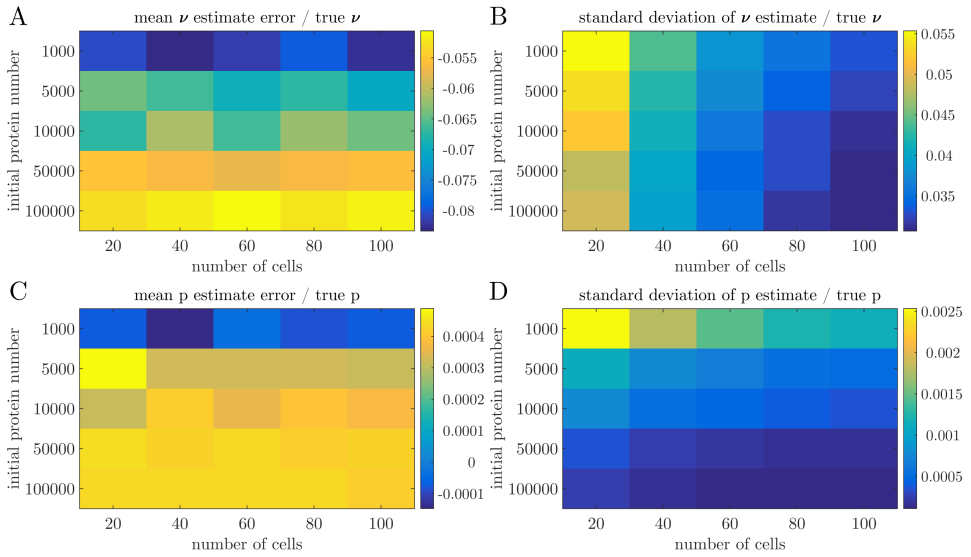

Figure 4.1: Performance of bleaching estimator on simulated data with no measurement error. Data was simulated using various numbers of cells and initial protein as shown on the x and y axis of each plot. p was kept at a constant value of 0.5 and ν at 10. In each case 1000 experiments were simulated and ν and p were estimated according to equations 4.7 and 4.6 for each experiment. The normalised difference between mean estimated parameter values and true parameter values is shown as a heat map in A and C with colour bars provided. In A yellow squares indicate more accurate estimation of ν , whereas in C the blue squares indicate a more accurate estimate of p. The normalised standard deviation in estimated parameter values across the 1000 experiments is plotted for ν and p in B and D, with blue squares corresponding to lower standard deviations. It can be seen that both ν and p are well estimated even for small numbers of cells.

Nayak and Rutenberg, but we have derived it from Bayesian principals making no assumption about the value of p.

The expression for p_{modal} proved difficult to work with analytically, but if we assume p_{modal} calculated from the data to be the true value of p we are able to calculate the expected value of ν_{modal} to be the true value ν_{true} (see appendix D). This shows that if an accurate value of p an be obtained from the data then an accurate value of ν will be obtained, provided we average our estimate over a sufficiently large number of cells. This matches the result of Nayak and Rutenberg and was confirmed by simulation results shown in figure 4.1. Panel A shows that the estimate of ν is always negative, and more so for lower protein numbers. This is similar to a systematic error which was see by Nayak and Rutenberg in their analysis. Even so, it can be seen that for even small numbers of cells ν can be accurately inferred.

4.2 Estimation of ν with Measurement Error

As discussed in earlier chapters there are many sources of error that affect the measurements from our microscope, and ignoring these is likely to lead to significant errors in our estimation of ν. Additionally, the simple context of a first order decay process allows us to infer the measurement error of our system in the same way we have inferred ν [51, 149], providing an opportunity to corroborate some of our conclusions about measurement errors from earlier chapters. To include measurement error we followed the work of Rosenfeld et al. [149] and posit a 'true' fluorescence measurement y_i. This obeys the conditional probability derived earlier and given in equation 4.4, that is:

$$p(y_{i+1}|y_i) = \phi_N(y_{i+1},\, y_i p,\, \nu y_i(1-p)p)$$

Here we have used the abbreviation $\phi_N(x, m, \sigma^2)$ for the probability density function of the normal distribution, with mean m and variance σ^2 assessed at the point x. The oberved values I_i are related to the true values y_i by an error distribution, $p\left(I_i|y_i, \Phi_M\right)$, with parameters Φ_M. We initially assume the error function is also normally distributed. Giving:

$$p(I_i|y_i, \sigma_e) = \phi_N(I_i, y_i, \sigma_e^2) \tag{4.8}$$

where σ_e^2 is the variance of the error. Combining these gives a final posterior of:

$$p(\nu, p, \sigma_e, \{y_0, ...y_M\}|\{I_0, ...I_M\}) \propto \prod_{i=0}^{M-1} \phi_N(y_{i+1}, y_i p, \nu y_i(1-p)p) \\ \phi_N(I_i, y_i, \sigma_e^2) p(\nu, p, \sigma, \{y_0, ...y_M\}) \tag{4.9}$$

The true fluorescence is of little interest and so we wish to marginalize over the set of variable $\{y_0, ...y_M\}$:

$$p(\nu, p, \sigma_e|\{I_0, ...I_M\}) \propto \int_{Y_0} \int_{Y_1} \cdots \int_{Y_M} \prod_{i=0}^{M-1} \phi_N(y_{i+1},\, y_i p,\, \nu\, y_i\,(1-p)\,p) \\ \phi_N(I_i,\, y_i,\, \sigma_e^2) p(\nu,\, p,\, \sigma,\, \{y_0, ...y_M\})\, d^{M+1}y \tag{4.10}$$

This integration could not be done analytically, and is computationally intensive if done in a naive numerical way. Following the example of Rosenfeld et al. [149], we implemented a variable elimination algorithm [198] to compute the numerical integration efficiently. Results from simulated data are given in figure 4.2 and in more detail in appendix D. These show that under reasonable experimental conditions, ν and σ_e can be accurately inferred by this procedure.

Though computationally intensive, the analysis described performs well on simulated data and can be straightforwardly adapted to other noise models such as

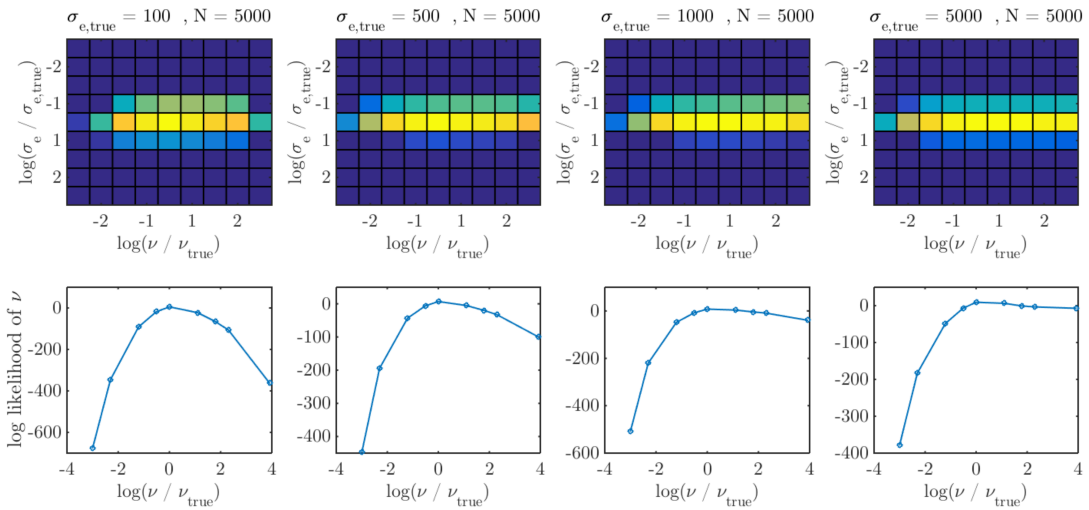

Figure 4.2: Performance of Bayesian estimator on simulated data with Gaussian noise. Data was simulated for cells undergoing a binomial degradation process with the addition of Gaussian noise. Specifically, 30 cells were simulated in each dataset for 30 time points with the parameters $\nu = 10, p = 0.5$ and an average of 5000 initial proteins. Gaussian noise was added to each data set with a standard deviation as shown in the title of each plot. For scale, the standard deviation expected from stochasticity would be approximately 4000 AU at the beginning of each simulation. The likelihood given by equation 4.10 was calculated over a grid of values in ν, σ_e and p for each simulated cell individually. The upper plot is a heat map showing the logarithm of the product of these individual likelihoods marginalised over p. The lower plot shows the same likelihood, additionally marginalised over σ_e to give the unnormalised posterior of ν . One sees that σ_e is well determined in each case, and that though the likelihood always has the correct modal value of ν , it broadens as σ_e increases. This is to be expected, since at higher σ_e the variations due to finite protein numbers are swamped. p was so well determined by the data that over the grid of p values used, no value other than the true one had a detectable likelihood. Estimates based on the noise free estimator given by equation 4.7 were completely inaccurate (see appendix D for details)

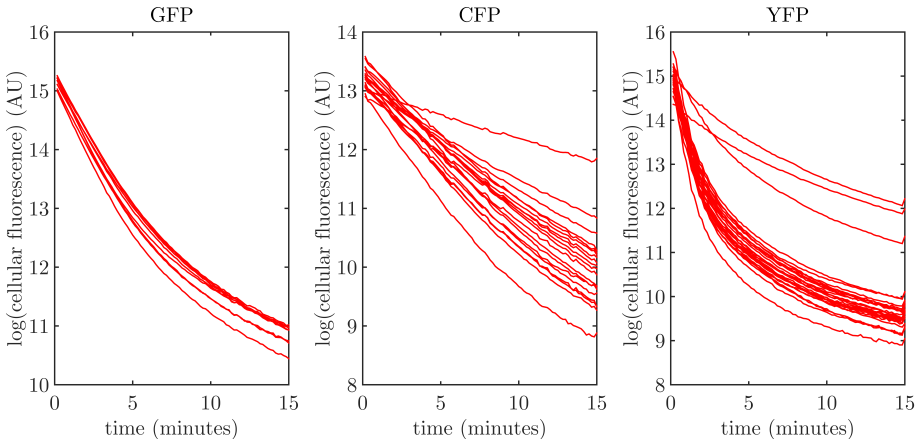

Figure 4.3: Bleaching data from highly expressing fixed cells. As explained in the main text, cells expressing high concentrations of CFP, YFP and GFP were fixed according to standard paraformaldhyde protocol and subjected to photobleaching by continuous illumination on an epifluorescence microscope. Images were obtained every 10 seconds over 15 minutes, and protein concentration estimated as total cellular fluorescence for each cell. The three plots show the logarithm of protein fluorescence plotted against time for the three different fluorophores. It can clearly be seen that in all cases the traces do not follow an exponential decay.

Poissonian or log-normal. We now describe the application of this procedure to photobleaching data acquired on our microscope.

4.3 Application to Data

To test whether these estimators could be applied to real data, we obtained photobleaching time series for fixed cells on our epifluorescence microscope(see appendix D for details). High expression of GFP, CFP and YFP was induced from the *GAL1* promoter and cells were fixed using a standard paraformaldhyde protocol (protocol in appendix B). This ensured that no additional protein would be produced during the experiment, which would have disrupted the analysis [124]. Cells were mounted on cover slips using concanavalin A and bleached by continuous exposure to excitory light of the wavelength appropriate to the fluorophore, while images were taken every 10 seconds. The results are shown in

figure 4.3 as plots of the logarithm of protein fluorescence against time. Clearly the traces do not follow an exponential decay, so the photobleaching of the fluorophores cannot be occurring via a first order process as we have assumed in the construction of our Bayesian estimator. A biexponential model with two populations decaying at different rates provides a much better description of the data.

Photobleaching is a complex process which, due to its implications for advanced imaging techniques such as Förster resonance energy transfer (FRET) and fluorescence lifetime imaging microscopy(FLIM) [176], has been comprehensively studied. It occurs due to chemical reactions of the excited fluorophore [41], and non-exponential fluorescence decay is commonly observed for a variety of reasons[53, 61, 92, 134]. If the fluorophore is at low concentrations then the predominant reaction partner is reactive oxygen, in which case photobleaching can occur via a pseudo first order process. However, if concentration is high then bimolecular fluorophore-fluorophore reactions can be a significant decay pathway, resulting in non-exponential behaviour [147, 165]. Decay rate is dependent on the intensity of excitory light, so uneven illumination across the field of view can lead to different decay rates between cells and, if sufficiently inconsistent across the cell, to multi-exponential photobleaching traces[12].

The micro environment of the fluorophore also determines decay rate, and can result in multi-exponential behaviour if the cell is not well mixed and individual molecules are in different chemical conditions. This can occur in the cell if fluorophores are located in various subcellular compartments or are bound to different molecules [164]. A similar effect is observed if dyes are not able to rotate freely, since the polarisation of the dyes relative to the excitory light can affect bleaching kinetics [53]. Reversible photobleaching, where fluorophores lose fluorescence only temporarily, is another phenomenon that could result in non-exponential behaviour. Reversible photobleaching is observed in many fluorescent proteins

and can dominate permanent photobleaching in GFP, YFP and CFP [157].

Given the slow variations observed in flat field measurements of chapter 2, and that we are using an epifluorescence microscope, it seems unlikely that there would be significantly inhomogeneous illumination in either the lateral or the vertical directions. GFP expressing cells were tested for reversible photobleaching by imaging after an extended period of excitation, and no evidence for reversible photobleaching found. Given the brightness of the cells it is seemed likely that the fluorophores are present at sufficiently high concentrations that the occurrence of protein-protein interactions was a plausible explanation for the non-exponential behaviour. We therefore looked at low expressed proteins in the hope that these would display a mono-exponential decay.

We focused on GFP, and in particular on the Hog1p-GFP strain obtained from the GFP fusion library [59], in our experiments. Hog1 is an osmotic stress protein that occurs at approximately 7000 proteins per cell [58, 111], close to the lowest concentration that will give a robust signal above background fluorescence. It is located in the cytoplasm and translocates to the nucleus under osmotic stress [24], so cells fixed in normal conditions should have a reasonably uniform distribution of cytoplasmic GFP.

Data for Hog1p-GFP cells is shown in figure 4.4. We observed that the traces still deviated significantly from a mono-exponential decay, and hypothesised that for these dimly expressing cells this may be due to autofluorescence. Autofluorescence is the commonly observed phenomena that cells and media are fluorescent even in the absence of fluorescent protein [14]. In earlier work within the lab, Lichten et al. found linear demixing [98] could effectively correct for autofluorescence in population level measurements made using a platereader [106]. In this procedure, measurements in the different fluorescent channels are assumed to be a linear transformation of the concentrations of the underlying fluorescent species. If the species have distinguishable spectra this transformation can be

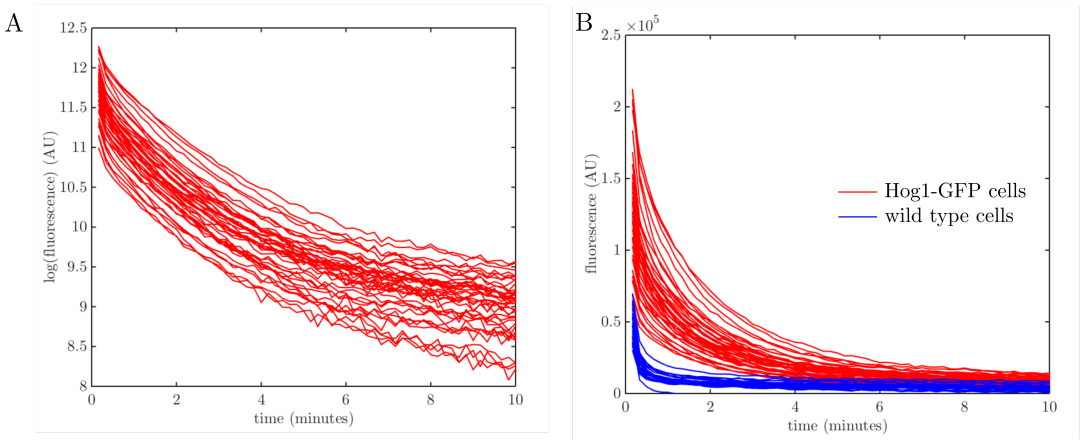

Figure 4.4: Photobleaching traces from fixed Hog1p-GFP cells. Cells were fixed using a standard paraformaldhyde protocol and mounted on microscope slides treated with concanavalin A. Photobleaching was performed over a period of 20 minutes by continuous illumination with excitory light, with images being acquired every 10 seconds, which were then processed by custom Matlab scripts (Full details are given in appendix D). Panel A shows natural logarithm of cellular fluorescence. It is clear that the traces do not obey a mono-exponential decay throughout the time course. Panel B shows the raw traces for the same cells in red, with fluorescence for wild type cells expressing no GFP in blue. The comparable fluorescence led to the hypothesis that, for these dim cells, a significant proportion of the non exponential behaviour could be due to autofluorescence - at least at the early time points.

reversed, demixing the signals. To apply this to autofluorescence, Lichten et al. imaged the samples twice at each time point: once with the conventional GFP filter sets, and once with standard excitation filters but an mCherry emission filter centred at 585 nm. The broader spectrum of autofluorescence meant that it had a proportionally larger signal in the second measurement, and this allowed the GFP signal to be distinguished from the autofluorescence. I will now discuss our attempts to apply the same procedure to single cell microscopy experiments.

4.3.1 Autofluorescence Correction by Linear Demixing

In order to apply the demixing procedure, it is necessary to have a second measurement with either different emission or excitation filters that provides information about a different part of the spectrum. Analogously to Lichten et al. [106] we adapted our microscope control software to image cells in a new channel, named GFPAutoFL, which is characterised by the same excitation filter as the GFP channel but uses a long wavelength tdTomato emission filter (605/70 nm). Given the broad spectrum of autofluorescence [106], this channel was expected to have a proportionally more significant contribution from autofluorescence than from GFP. The microscope was configured to image with both these filter sets every 10 seconds while cells were continuously illuminated with GFP excitation light to induce photobleaching. A brightfield image was obtained at the end of the experiment for cell identification and segmentation.

Analysis With Additional GFPAutoFL Channel

We label measurements from the normal GFP filter sets I^g and those from the new GFPAutoFL channel I^a. We adopt the standard linear [98] mixing model

in which these measurements are related to the GFP (y^G) and autofluorescence (y^A) signals by the equation:

$$\begin{pmatrix} I^g \\ I^a \end{pmatrix} = \begin{pmatrix} \alpha_{G,g} & \alpha_{A,g} \\ \alpha_{G,a} & \alpha_{A,a} \end{pmatrix} \begin{pmatrix} y^G \\ y^A \end{pmatrix} + \begin{pmatrix} \beta_g \\ \beta_a \end{pmatrix} \tag{4.11}$$

Here α is the mixing matrix and the vector β is a flat offset that can be different for each channel. Given that the measurements are subject to error, this is better written as:

$$p\left(I^g, I^a | y^G, y^A, \Phi_M\right) \sim N\left(\begin{pmatrix} \alpha_{G,g} & \alpha_{A,g} \\ \alpha_{G,a} & \alpha_{A,a} \end{pmatrix} \begin{pmatrix} y^G \\ y^A \end{pmatrix} + \begin{pmatrix} \beta_g \\ \beta_a \end{pmatrix}, \begin{pmatrix} \sigma^2_{e,g} & 0 \\ 0 & \sigma^2_{e,a} \end{pmatrix}\right) \tag{4.12}$$

where Φ_M is the set of all the measurement parameters (i.e. the α's, β's and σ's present in the above equation) and N is the normal distribution.

From a Bayesian perspective we would now ideally construct a model based likelihood using the whole data set $\{I^g_0, I^a_0, ...I^g_M, I^a_M\}$, and then marginalise over both y^A and y^G to obtain a likelihood for the parameters of interest. This would require an explicit model of autofluorescence, which is not obvious, and an additional set of computationally intensive numerical integrals. Instead, we applied the commonly used deterministic demixing algorithm to remove the autofluorescent contribution.

By rearranging the equations for I^g and I^a we can define a new quantity, I^s, from

which we have eliminated the contribution of y^A:

$$\begin{aligned} I^s &= (I^g - \beta_g) - \frac{\alpha_{A,g}}{\alpha_{A,a}}(I^a - \beta_a) \\ &= I^g - r_A I^a + c \end{aligned} \tag{4.13}$$

This requires that the mixing matrix α is not unitary, which corresponds to the intuitive idea that the two spectra cannot have proportional contributions.
Given the distribution of I^g and I^a, we can see that:

$$I^s \sim y_G \left(\alpha_{G,g} - \frac{\alpha_{G,g}\alpha_{A,g}}{\alpha_{A,a}} \right) + N(0, \sigma_g) + N(0, \sigma_a)\frac{\alpha_{A,g}}{\alpha_{A,a}}$$

We see that I^s has no contribution from y^A and is normally distributed around y^G, allowing us to apply the Bayesian fitting procedure described in section 4.2 to data obtained by linear demixing.

This procedure was applied to data obtained from our microscope. Both Hog1p-GFP and wild type cells were bleached according to the standard protocol (see appendix D), acquiring both GFP and GFPAutoFL channels. Equation 4.11 and the analysis derived from it assume a linear relationship between these two channels from fluorescence produced by both the GFP protein and autofluorescence, which we attempted to confirm using highly expressing and wild type cells. To confirm the linear relationship between the GFP and GFPAutoFL measurements for fluorescence coming from the GFP protein was assessed using data from strongly induced Gal1-GFP cells, while the autofluorescence correction parameters (r_A and c) were calculated at each time point of the acquisition from linear fits of GFP and GFPAutoFL measurements for wild type cells. This is shown in panels A and B of figure 4.5.
The unnormalised posterior of the parameters (ν, p, σ_e) given this data was sam-

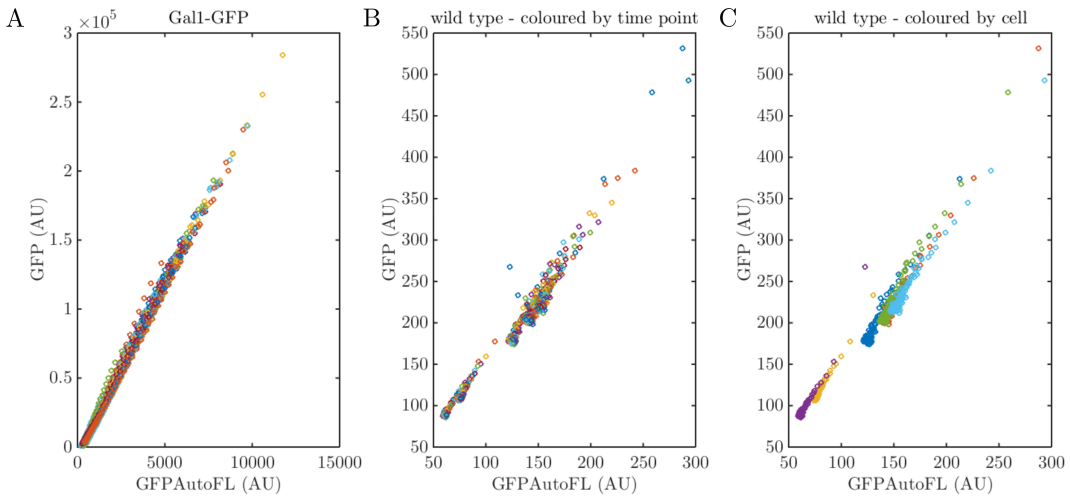

Figure 4.5: Ratio of GFP to GFPAutoFL emission for highly expressing and WT cells. The demixing analysis, formalised in equation 4.11, assumes a linear relationship between these two channels for fluorescence produced by both the GFP protein and autofluorescence. We investigated the validity of this assumption using highly expressing and wild type cells. Strongly induced *GAL1*-GFP cells and wild type cells were imaged with both the GFP and GFPAutoFL filter sets over a whole photobleaching experiment. Panel A shows GFP channel measurements plotted against GFPAutoFL for *GAL1*-GFP expressing cells. Panels B and C show the same data but for wild type cells. In Panel B each colour denotes a different time point, while in C a different cell. Panel C shows that the autofluorescent spectra varies from one cell to the next, confounding a simple demixing procedure.

pled using an adaptive Markov chain Monte Carlo (MCMC) scheme [68] and the likelihood defined by equation 4.10. The mean of ν from this sample gave an average cellular protein content of approximately 10 proteins, well below the 7000 expected [59] and certainly erroneous.

4.3.2 Improved Error Model

A possible reason for the poor performance could be an inaccurate error model. As discussed in chapter 2 Poisson noise, or shot noise, from the camera can make a significant contribution to measurement error. We hypothesised that since this would not be well modelled by the identically distributed Gaussian error assumed in our analysis, it may contribute to the poor performance observed. To correct for this, we used the camera noise measurements from 2 to estimate the variance of the Poisson noise affecting each pixel constituting the cell. These were adjusted to account for multiplication by the flat field correction and summed for each cell at each time point individually. This gave an estimate of the shot noise contribution to measurement error for each cell at each time point. Errors were similarly calculated for background correction and propagated through the extraction.

The distribution of errors so calculated is shown in figure 4.6. Since the final value of cellular fluorescence is produced by subtracting a noisy background measurement, propagation of errors leads to a variance that is higher than would be expected from pure Poisson noise. A linear fit gives a ratio of around 1.5 between variance and measured fluorescence.

This variance estimate can be straightforwardly integrated into our error function. If we approximate it by a Gaussian, which is reasonable for a Poisson distribution

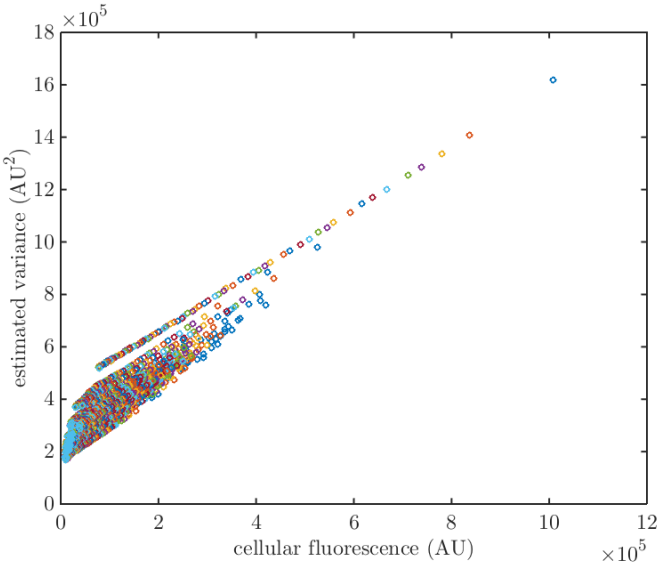

Figure 4.6: Estimated variance from camera noise plotted again total cellular fluorescence. As described in the main text, the variance is calculated from the cameras variance-intensity relationship measured in chapter 2. Since the final value of cellular fluorescence is produced by subtracting a noisy background measurement the variance is higher than would be expected from pure Poisson noise, and a linear fit gives a ratio of approximately 1.5 between variance and measured fluorescence. This fit is only for guidance, the raw camera noise estimates plotted being used in the analysis.

with high intensity, then our measurement error function becomes:

$$p(I_i, \sigma_{\text{est},i}|y_i, \sigma_{\text{shot},i}, \sigma_e) = \phi_N(I_i, y_i, \sigma_e^2 + \sigma_{\text{shot},i}^2)p\left(\sigma_{\text{est},i}|\sigma_{\text{shot},i}\right)$$

where $\sigma_{\text{shot},i}$ is the true camera noise and $\sigma_{\text{est},i}$ is the estimate we have made from I_i. Assuming that this estimate is accurate (equivalent to assuming $p\left(\sigma_{\text{est},i}|\sigma_{\text{shot},i}\right)$ to be a delta function) and marginalising over $\sigma_{\text{shot},i}$ we obtain.

$$p(I_i, \sigma_{\text{est},i}|y_i, \sigma_e) = \phi_N(I_i, y_i, \sigma_e^2 + \sigma_{\text{est},i}^2) \tag{4.14}$$

This is obviously a strong assumption, but it allows us to avoid imposing a restrictive or complex model for the estimated camera error. Inspection of the traces shows it to be a relatively small contribution to the total variance.

4.3.3 Autofluorescence Correction by Whole Sample Subtraction

Looking at more detail at the data of individual wild type cells we saw that the parameters r_A and c seemed to vary significantly across the population. This can be seen in panel C of figure 4.5, and was confirmed by applying the correction obtained from one set of wild type cells to another. The result is shown for a few cells in panel A of figure 4.7, where it can be seen that the difference in autofluorescence parameters lead to a significant systematic error.

An additional problem is that I^a, being a lower signal, is more noisy than I^g. When it is multiplied by r_A and subtracted from I^g it introduces substantial unwanted noise. This can be seen from equation 4.3.1 and panel B of figure 4.7. Both these factors may explain why demixing did not result in an improved inference result, and led us to search for another means of dealing with autofluo-

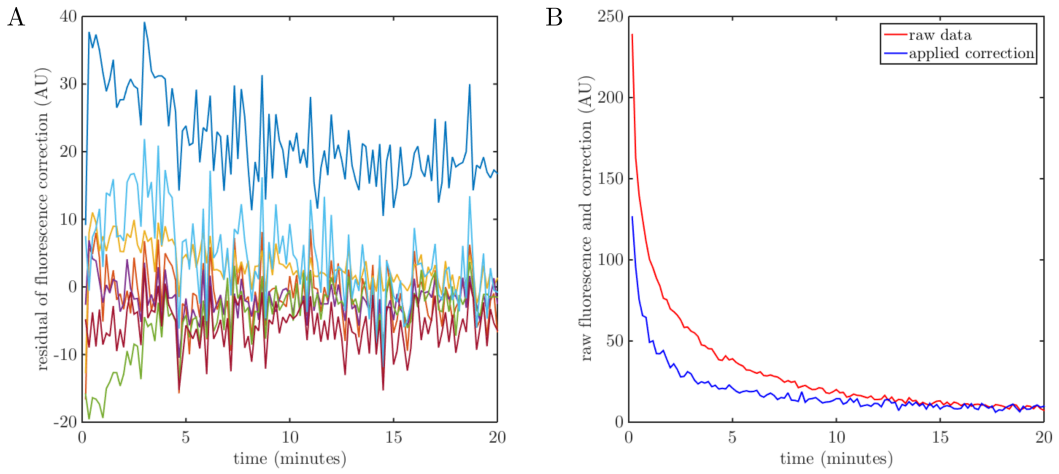

Figure 4.7: Result of autofluorescence correction using population averaged spectral parameters. In panel A demixing is applied to one set of wild type cells using autofluorescence parameters found from another. If the correction were successful the result should be Gaussian noise around an average value of 0, and it can be seen that there is still a significant systematic error. Panel B shows the trace from a single Hog1p-GFP cell and the correction that would be subtracted if demixing would be applied. It can be seen that the correction introduces significant additional noise.

rescence.

Ignoring the GFPAutoFL channel, we recast the problem as follows. We begin as in equation 4.9 with the likelihood for the parameters and the true fluorescence $\{yi\}$, but extend it to include a putative true autofluorescence $\{a_i\}$:

$$p\left(\{y_i, a_i\}, \sigma_e, \nu, p | \{I_i, \sigma_{\text{est},i}\}\right) = p\left(\{y_i, a_i\} | \nu, p\right) \phi_N(I_i, y_i + a_i, \sigma_e^2 + \sigma_{\text{est},i}^2) \quad (4.15)$$

where $\{y_i, a_i\}$ is a set of true fluorescence and autofluorescence values and we have again assumed Gaussian measurement error.

Assuming that a does not depend on the fluorescence parameters, and again

marginalising over the variables $\{y_i, a_i\}$ we obtain:

$$p\left(\sigma_e, \nu, p|\{I_i, \sigma_{\text{est},i}\}\right) = \int_{Y_0}\int_{Y_1}\cdots\int_{Y_M}\int_{A} p\left(\{y_i\}|\nu, p\right) p(\{a_i\}) \prod_{i=0}^{M-1} \phi_N(I_i - a_i, y_i, \sigma_e^2 + \sigma_{\text{est},i}^2)\, da\, d^M y \tag{4.16}$$

where the integral over a is an integral over all possible paths $\{a_i\}$ and we have used the properties of Gaussians to rearrange the error term.

We do not know the form of $p(\{a_i\})$ to do the integral, but our measurement of wild type cells constitute a large number of noisy samples from $p(\{a_i\})$. We can use these to approximate the integral by sampling, giving:

$$p\left(\sigma_e, \nu, p|\{I_i, \sigma_{\text{est},i}\}\right) \approx \int_{Y_0}\int_{Y_1}\cdots\int_{Y_M} \sum_{a^j \in \{a\}_{\text{WT}}} p\left(\{y_i\}|\nu, p\right) \prod_{i=0}^{M-1} \phi_N(I_i - a_i^j, y_i, \sigma_e^2 + \sigma_{\text{est},i}^2)\, d^M y \tag{4.17}$$

Practically, what this corresponds to is subtracting all the wild type traces from the fluorescence trace for each Hog1p-GFP cell to produce a collection of W traces, where W is the number of wild type cells measured. The likelihood is calculated for each of these and summed to give a total likelihood.

Since the posterior for p is very well confined, we adapted the analysis by determining p by a least squares fit of a straight line to the logarithm of the data, only calculating the likelihood over a grid in ν and σ_e. Since we now have a collection of traces for each cell, we fit p by a straight line to the logarithm of each possible trace and take an average of these values weighted by their least squares fit scores. Results from simulations are shown in figure 4.8. It can be seen that the subtraction of wild type traces improves the performance, and with sufficient protein and reasonably low noise, ν can be faithfully estimated.

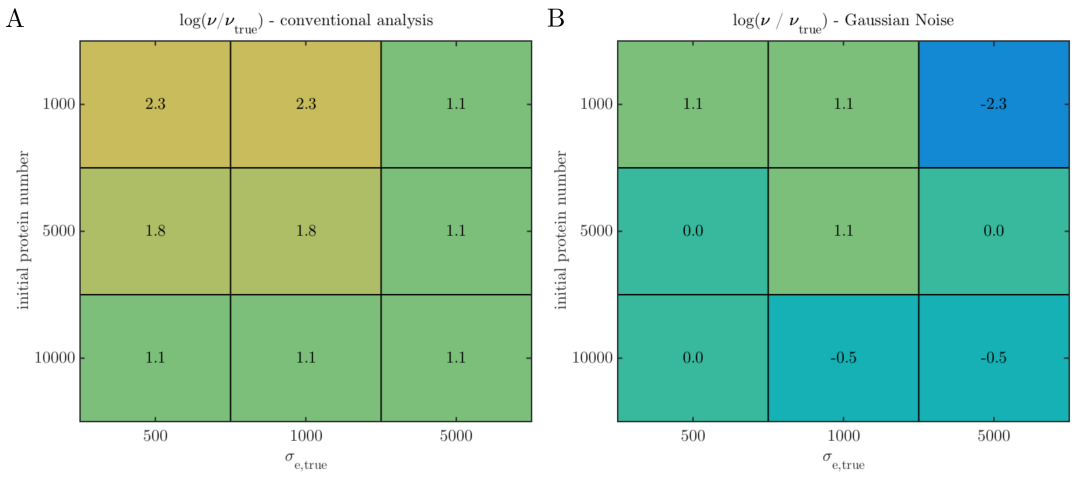

Figure 4.8: Subtraction of WT traces according to equation 4.17 improves the estimation ν on simulated data. Decay traces were simulated for 20 cells from an exponential decay with a ν of 30, p of 0.95 and a range of initial protein numbers and Gaussian noise intensities. This data was added to 20 wild type traces obtained from a bleaching experiment, while a further 40 were kept as a sample from the distribution of autofluorescent traces. Both Panels show the natural logarithm of the $\nu_{estimated}/\nu_{true}$ for various starting protein numbers (y axis) and Gaussian noise standard deviations (x axis), where $\nu_{estimated}$ is the maximum a posteriori estimate of ν over a grid of values. 0 indicates perfect agreement while positive and negative values indicate over and underestimation respectively. Panel A shows the result for the naive likelihood, assuming no autofluorescence, and given by equation 4.10, while panel B shows the improved result when using the likelihood given by equation 4.17. As described in the main text, the parameter p was estimated by least squares fit to the data.

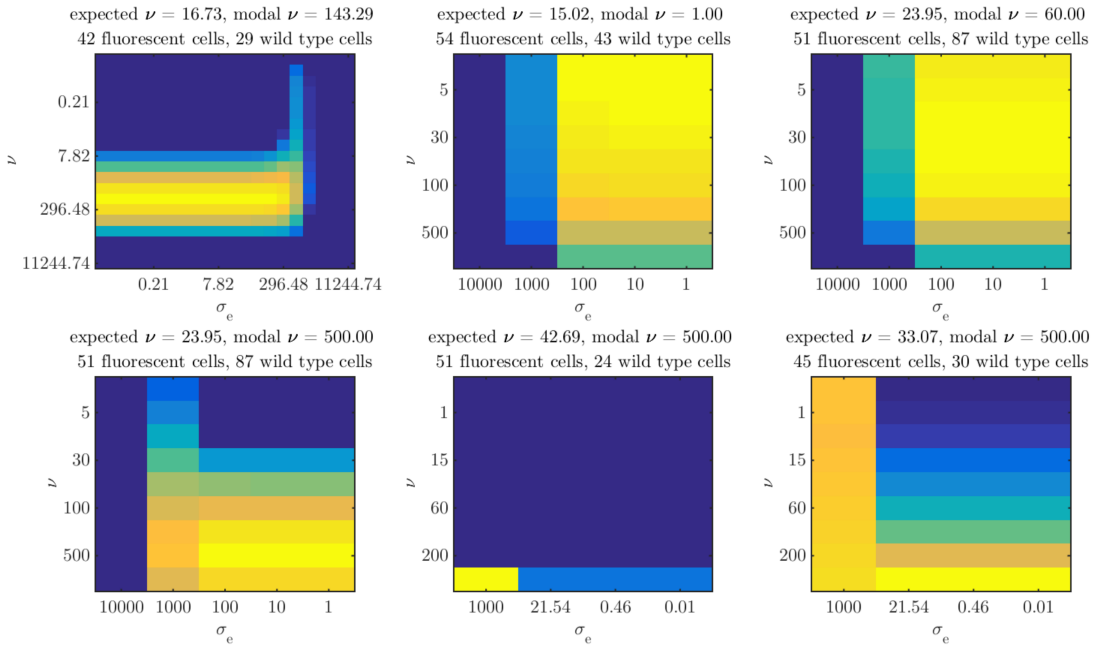

Figure 4.9: Results of parameter inference after applying autofluorescence correction by whole sample subtraction. In each case a set of fixed Hog1p-GFP and wild type cells were photobleached and the the approximate posterior given by equation 4.17 calculated over a grid of values for ν and σ_e , p being fixed at a value determined by least squares fitting. In all cases only 40 time points (5 minutes) was used, selected to avoid early non-exponential behaviour and the later period in which autofluorescence dominates. Heat maps show log likelihood at each of the grid of values, with an expected ν for the dataset based on the initial fluorescence measurement of the cell and literature value of 7000 Hog1p proteins per cell [59]. Though the applied analysis is able to bound σ_e well the modal estimates of ν are inconsistent and generally far from the value expected.

This estimator was applied to a collection of real data sets for Hog1p-GFP cells fixed and bleached as previously described: results are shown in figure 4.9. It can be seen that the estimated ν is not consistent between experiments. An expected value for ν was calculated for each experiment using the initial fluorescence of the cells and the literature value of 7000 Hog1 proteins [59]. This is shown in the title of each plot, and is far from the value estimated by our procedure. σ_e is also not well confined, with only a consistent upper bound.

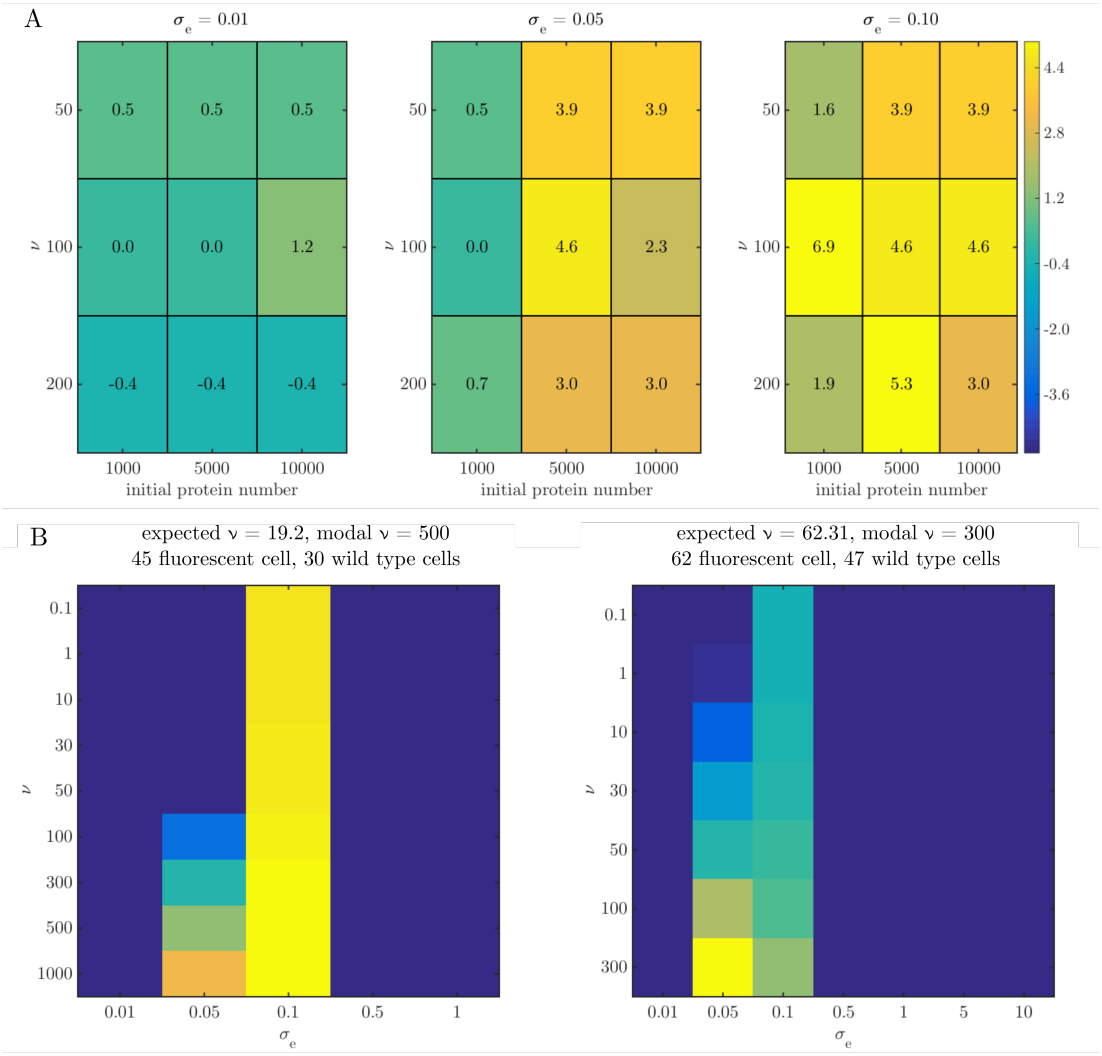

Figure 4.10: Results from simulation and actual data assuming log normal noise. The analysis was adapted to assume log normal noise and applied to both simulations of decay traces with log normal noise and to actual data. Simulations were performed as previously described: a data set for wild type cells was taken and simulated data added to a subset of the traces, while the rest were kept for autofluorescence correction. The results for analysis applied to this simulated data is shown in panel A. The colour of the heat maps indicates the logarithm of the ratio of ν_{true} and ν_{est} (the maximum likelihood value of ν) for simulations run with different values of ν and σ_e . Text is the value of the log ratio, with 0 indicating perfect agreement and positive and negative values indicating over and under estimation of ν respectively. The results show that for low values of σ_e , ν can still be accurately inferred, while σ_e is always accurately estimated (data not shown).
Panel B shows the log likelihood calculated by applying the same procedure to two sets of Hog1p-GFP photobleaching data. In these plots the brightness of the heatmap indicates the likelihood for the given set of parameters; the x and y axis are the tested values of σ_e and ν respectively. In both cases the estimated error is reasonably large, while ν is far from the expected value.

4.3.4 Adapting the Estimator to Log Normal Noise

It has been observed in the literature that for single cell fluorescence experiments, log normal noise may be more appropriate model than normal noise [194]. It is defined by the distribution:

$$\log(y) \sim N(\mu, \sigma) \tag{4.18}$$

and has the property that the standard deviation is approximately proportional to the mean. This may be an appropriate model if errors in the experiment are due to illumination, segmentation or changes in focal position.

Our analysis was straightforwardly adapted to log normal noise by simply changing measurement error distribution to:

$$p(\log(I_i)|y_i, \sigma_e) = \phi_N(\log(I_i), \log(y_i), \sigma_e^2 + \sigma_{\text{est},i}^2) \tag{4.19}$$

Results from the application of this modified analysis to real and simulated data are shown in figure 4.10. In the case of Hog1p-GFP the modal value of σ_e is reasonably large, and that of ν is far from the expected value. Looking at the simulated results, we see that ν is poorly estimated when σ_e is large, which may mean that our log normal noise model is correct but that our data is not sufficiently informative to accurately infer ν . Further work would be required to establish if this is the case, or if this unphysical ν is the result of an inappropriate model for our data.

4.4 Discussion

In this chapter I have detailed our efforts to combine experimental and analytical techniques to infer important experimental parameters from measurements of

cells undergoing photobleaching. We derived analytical results for the case without measurement error, and implemented an efficient computation of a Bayesian likelihood accounting for measurement error. Though straightforward in theory application to actual data has proven difficult. Despite numerous adaptations the result is so far unsatisfying, but investigating this problem has led to a far deeper understanding of our experimental system. It has motivated much of the characterisation work of chapter 2, while intermediary results such as the autofluorescence correction may find application in other projects in the future.

In order to use photobleaching to successfully characterise our microscope a number of problems have to be overcome. Currently the slow speed of both the experiments and the analysis prevent thorough characterisation and impede progress. Computational speed might be improved by using more sophisticated techniques, such as those specifically for stochastic time series analysis [51, 191]. This might also facilitate extending the underlying biophysical model to a bi-exponential decay, which appears to be a more appropriate model for most of the expression regime. Our efforts have focussed on finding an experimental regime where assuming mono-exponential decay is reasonable, but if our analysis was applicable to a bi-exponential decay then we could apply it to highly expressed proteins. This would avoid the complications of autofluorescence, which would simplify the experimental procedure substantially. Currently our background and autofluorescent corrections necessitate four separate photobleaching experiments - fluorescent cells, wild type cells and two background corrections - to be performed to obtain a single data set. Most of these could be dispensed with if cells were well above the level of autofluorescence and background.

Extension to bi-exponential decay would require additional assumptions about the underlying model, but if the result was faster analysis and experiments these assumptions could be tested and could result in important insights into the biophysics of photobleaching in cells.

In conclusion, this attempt to apply Bayesian inference to a set of time series data has not been successful, but pursuing it has revealed a great deal to us about our experimental system and the importance of appropriate models in Bayesian analysis.

Chapter 5

Investigation of *GAL1* Transcriptional Regulation and the Role of the *GAL10*-lncRNA

5.1 The *GAL* Network in *Saccharomyces cerevisiae*

We now look at an application of our microfluidic system and associated analysis to understanding transcriptional regulation: specifically, to understanding the regulation of the galactokinase encoding gene *GAL1*.

The *GAL* system - the collection of genes involved in the transcriptional response to the monosaccharide galactose - has been intensively studied as a model system for gene regulation and cellular decision making [88, 139, 177]. It is generally thought to have three states determined by the predominant carbon source: repressed (occurring in high glucose concentration), induced (requiring both galactose and the absence of glucose) and uninduced (seen in raffinose or glycerol).

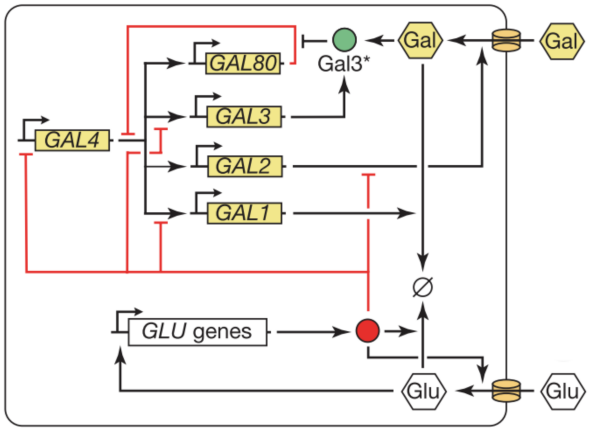

Figure 5.1: The figure depicts the major interactions of the *GAL* network in *Saccharomyces cerevisiae* . Activation effects are shown in black and repressive ones are shown in red. Though glucose is included its description here is simplified, showing only the significant points of repression. The figure is reproduced with permission from Bennett et al. [10].

Transcriptional activation is induced by Gal4p, which binds a well defined upstream activation sequence (UAS_g) found in the promoters of *GAL2*, *GAL7*, *GAL3* and the *GAL1-10* bidirectional promoter. In the absence of galactose Gal4p is bound by Gal80p, which represses its transcriptional activation effect, but when galactose is present an interaction between Gal3p and Gal80p relieves this repression. The precise mechanism for this reaction is still debated [1], but it seems that the formation of a complex of Gal80p, Gal3p and Gal4p is necessary. Later in the induction process Gal3p can be replaced by Gal1p, a paralogue of GAL3 [90] with weaker activation capabilities but much higher expression when the *GAL* system is fully induced [110].

The involvement of galactose, Gal3p and Gal80p in the induction of the system by which they are regulated leads to a number of positive and negative feedback loops which are depicted in figure 5.1. *GAL2*, *GAL1* and *GAL3* are all upregulated by the Gal4p-Gal3p-Gal80p mechanism described above. Gal1p and Gal3p relieve Gal80p repression in the presence of galactose which Gal2p imports, all leading to positive feedback loops. Antagonistic to this are two negative feedback loops: the metabolism of galactose by Gal1p, which reduces its concentration in the cell, and the upregulation of *GAL80* by the Gal4p-Gal3p-Gal80p mechanism. The interaction of these network motifs leads to interesting emergent phenomena such as bistability [2] and memory [168], which have been shown to depend on

both the network structure of the *GAL* system and the variability in its constituent proteins [2, 86, 93]. These properties, and the large amount of data and information available on the *GAL* network, have made it a popular subject for modelling studies which have helped to elucidate these network behaviours [3, 163].

In the *GAL* system glucose is the main antagonist to galactose, and its presence at high concentration can cause complete repression of the major *GAL* genes. This is effected mainly through three mechanisms: the active degradation [82] and inhibition of Gal2p and the repression of *GAL4* and *GAL1*. These last two require the protein Mig1p, which binds an upstream recognition sequence (URS) in the *GAL1-10* promoter [91].

The properties described above have been largely identified at high concentration of galactose and glucose. More recent work is finding that the interactions of multiple sugars at a greater range of concentrations reveals new and interesting phenomena. Escalante-Chong et al. [48] looked at the behaviour of populations exposed to mixtures of galactose and glucose across a wide range of concentrations, and found the proportion of cells inducing *GAL* genes was determined by the ratio of the two sugars. This observation held over several orders of magnitude of concentration and across different *Saccharomyces cerevisiae* strains, and could be explained by competitive binding to an as yet unidentified transporter. Using a microfluidic device and dynamic glucose-galactose environments, Bennett et al. [10] were able to show that the transcripts of a number of *GAL* genes were actively degraded in glucose, revealing that the response to glucose occurs at the level of transcription, mRNA degradation and targeted proteolysis [82]. As a last example, Houseley et al. [84] showed that a long non coding RNA (lncRNA) running antisense to the *GAL10* gene causes changes in the dynamics of *GAL1* transcription. This was only observed when both galactose and glucose were present at low concentrations. The *GAL10*-lncRNA is one of a number of

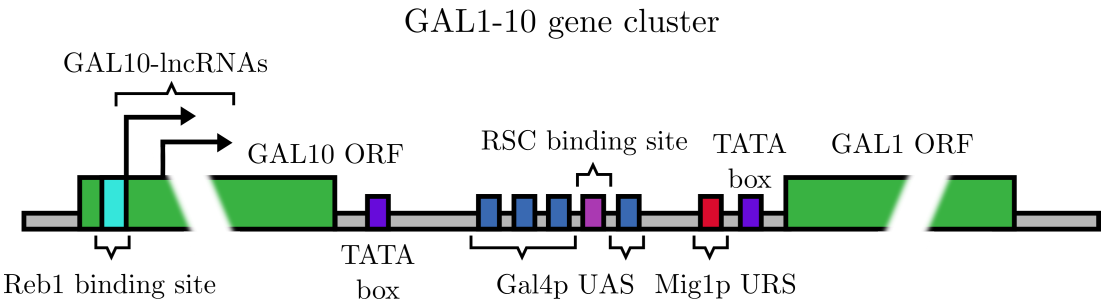

Figure 5.2: A depiction of major regulatory elements in the *GAL1-10* bidirectional promoter. The four Gal4 UAS_g sites are responsible for the galactose induced expression mediated by Gal4p. The RSC binding site maintains an open chromatin state, while the Mig1 URS and Reb1p activated lncRNAs contribute to glucose mediated repression. More details of the mechanism and effect of each element is given in the main text.

lncRNA in the *GAL* system, and is a subject we will return to in greater detail later in the chapter.

The introduction above shows the depth of knowledge available about the *GAL* system, illustrating why it is a canonical example of a gene network. The regulation of the individual genes by the interaction of different mechanisms has also been studied in detail, leading to insights applicable beyond *Saccharomyces cerevisiae* [177]. We next described the various mechanisms that have been found to function in the regulation of the *GAL1-10* bidirectional promoter, which will be the focus of the work in this chapter.

5.1.1 Regulation of *GAL1*

GAL1 and *GAL10* both encode metabolic enzymes in the *GAL* system and are transcribed from a bidirectional promoter. Both proteins are required for galactose metabolism [44], are highly expressed and tightly regulated [110], while *GAL1* expression is commonly used as a readout for activity of the whole *GAL* network [2, 10, 48, 100]. A diagram of the *GAL1-10* locus with some of its major regula-

tory elements is shown in figure 5.2.

As described above, the major transcriptional activator of the *GAL* network is Gal4p, which is relieved of Gal80p inhibition by either Gal3p or Gal1p in the presence of galactose. The *GAL1-10* promoter has 4 UAS_g to which Gal4p can bind, leading to strong induction in the presence of galactose. Additionally, the *GAL1-10* promoter has a binding site for the chromatin remodelling complex (RSC), which locally manipulates the nucleosomes to maintain a region of exposed DNA around the UAS_g s. Cells with this RSC binding site deleted show a reduced localisation of Gal4p to the promoter when transitioning from glucose to galactose, and correspondingly a slower induction of *GAL1* [52].

As well as depending on an actively maintained chromatin configuration at the *GAL1* promoter, Gal4p also modifies the chromatin via the recruitment of the nucleosome remodelling complex SWI/SNF. This occurs during induction and precipitates the removal of nucleosomes from the *GAL1-10* promoter, and in the absence of SWI/SNF both nucleosome removal and induction are slowed. Upon repression with glucose the chromatin is again deposited, but if galactose is present in the media this occurs only slowly. Despite this extended maintenance of the open chromatin state *GAL1* is still repressed, all other transcriptional machinery being removed from the gene [22]. The action of SWI/SNF has been shown to be important for reinduction memory but only in the short term. Long term memory is instead mediated by Gal1p concentration [100].

The nucleosomes constituting the chromatin at the *GAL1-10* locus are not of the normal variety, but contain a variant of the histone core protein H2A, H2A.Z, which is associated with the promoters of regulated genes [105]. Cells without the special H2A.Z histone show a slower and weaker induction of galactose [71], though of course the removal of H2A.Z is a global perturbation not restricted to the *GAL1-10* locus.

A further layer of regulation is the location of the *GAL1-10* locus within the

nucleus, which Cabal et al. [23] showed is preferentially located at the nuclear periphery during *GAL* induction. Fluorescence in situ hybridisation (FISH) of *GAL1* transcripts associated this nuclear periphery localisation with transcription, though abrogating peripheral localisation did not impede expression [23].
As described earlier, glucose repression of the *GAL* genes occurs by a number of mechanisms, one of which is the upstream repressive sequence (URS) specific to the *GAL1-10* promoter. This is bound by the Mig1 protein and causes an approximate 10 fold reduction in expression relative to a synthetic promoter with no URS [91]. An additional mechanism of glucose repression is the *GAL10*-lncRNA discussed earlier, but this appears to contribute a significant repressive effect only in mixes of glucose and galactose.

The work described above illustrates the strengths of the *GAL1* gene for our purposes. Its regulation involves the integration of many mechanisms that are well characterised, can have subtle influences on *GAL1* expression and have not been thoroughly investigated at the single cell level. These effects can be kinetic, and can persist over hours or generations, but their persistence within a single cell has not been assessed. It is our belief that with the large data sets obtainable from our microfluidic device, and the thorough understanding of our experimental system developed in preceding chapters, we will be able to identify and understand subtle changes arising from the genetic perturbation of the regulation mechanisms described above. Further, we hope that by applying Bayesian inference to the data so obtained we will not only be able to observe the effect of a particular control apparatus, but be also able to produce quantitative models of promoter behaviour under different environmental and genetic conditions.
As a first investigation and proof of principle we used our microfluidic system to look at the influence of the *GAL10*-lncRNA on the expression of *GAL1*. Specifically, we compared expression of a Gal1p-GFP fusion in wild type strains and

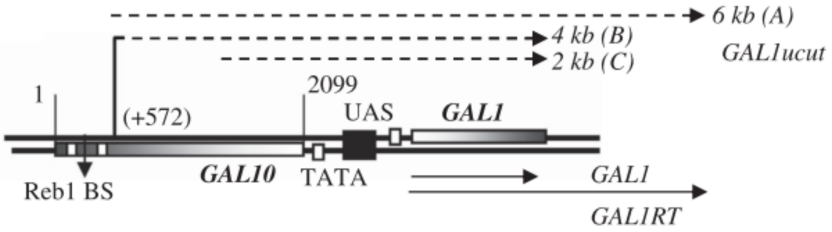

Figure 5.3: Location of the Gal10 lncRNAs. Gal10 has 3 lncRNAs all transcribed antisense to the Gal10 gene and overlapping the Gal1-10 promoter. The Reb1 binding sites related to the 4kb transcript are those removed in Houseley et al. [84]. Figure reproduced with permission from [140]

Reb1BSΔ in which the expression of the *GAL10*-lncRNA has been abrogated [84]. Since its discovery by Houseley et al., the *GAL10*-lncRNA has been the subject of a number of studies that have resulted in apparent discrepancies, both in the conditions under which the long non-coding RNA is active and its effect on *GAL1* expression. To try and give a complete picture of current understanding we will begin with a review of the major work on the *GAL10*-lncRNA.

5.1.2 Review of Work Pertaining the *GAL10*-lncRNA

Antisense long non-coding RNAs are a common form of gene regulation and have been discovered across the yeast genome. They can affect transcription initiation, elongation, RNA stability and chromatin state and are thought to be a fast evolving regulation system for controlling specific genes or loci [137].

There are 3 lncRNAs in the *GAL1-10* locus which all start within the *GAL10* coding sequence, and are transcribed antisense to the *GAL10* ORF. They were first reported in Houseley et al. [84], where they were identified by histone residue methylation patterns concomitant with transcription from the 3' end of the *GAL10* coding sequence. Reb1 binding sites were found in this region, and their abolition was seen to remove the unusual methylation patterns observed. Subsequently, RNA detection methods identified 3 lncRNAs (2.3, 4 and 5.6 kb) and

establish that the Reb1 binding sites removed corresponded to the start of the most abundant 4kb lncRNA. Further work showed that this lncRNA was transcribed in glucose and raffinose media but not galactose. Comparison of wild type (WT) and Reb1 binding site knock out cells (Reb1BSΔ) grown in 2% glucose or 2% galactose showed no significant differences in *GAL1/10* mRNA concentration. *GAL1* mRNA time courses for cells induced with 2% raffinose/ 2% galactose mix after overnight growth in 2% raffinose were also statistically indistinguishable Reasoning that cells were unlikely to experience such high and pure sugar concentrations in the wild, the authors looked for differences in *GAL1* and *GAL10* expression between WT and Reb1BSΔ strains at a range of low concentrations and mixtures of galactose and glucose. Induction with 2% raffinose/ 0.01% galactose/ 0.02% glucose mix gave reproducible and statistically significant differences, and this condition was subsequently used in a time series induction experiment. This showed significant differences in *GAL1* expression kinetics between the two strains over the first 6 hours of induction, establishing a role for the *GAL10*-lncRNA in *GAL1* regulation.

Studies of heterozygous diploids showed that the lncRNA acted only in *cis*, which together with the earlier measurements of histone methylation patterns led the authors to hypothesise that the lncRNA functioned via methylation mediated recruitment of a histone deacetylases (HDAC). This was confirmed, and recruitment of the Rpd3S complex found to be the mechanism of repression.

Pinskaya et al. [140] took a different approach that was based on global perturbation rather than the mutation of the specific Reb1 binding sites in the *GAL10* 3' region. The authors measured *GAL1* mRNA concentration 1 hour after induction with 2% galactose in a range of strains and found that *set1*Δ showed a significantly increased induction, while *set2*Δ was indistinguishable from WT. Given the role of Set1p in establishing H3K4 methylation this observations pointed to a role for H3K4 methylation in *GAL1* repression, which was confirmed by sim-

ilar experiments on H3K4A strains [1]. Measurements of mRNA stability and pol II occupancy at the *GAL1* locus showed that expression changes were due to transcription initiation rather than mRNA stability. The same lncRNAs were identified as in Houseley et al. [84] and the role of Reb1 also confirmed by temperature sensitive Reb1 degradation strains (*reb1-1*). The lncRNAs were stabilised by deletion of *XRN1* and *TRF4*. Though this was said to have no effect on *GAL1* induction the data for this statement are limited, and as we will see conflicts with other work [34]. Measurements of H3K4 tri-methylation across the locus in different strains and conditions showed that while *set1*Δ and WT strain are significantly different in both glucose and galactose, *reb1-1* and *set1*Δ tri-methylation patterns are equivalent in glucose, indicating that Reb1p and Set1p act synergistically to alter H3K4 methylation.

Cloutier et al. [34] investigated the effect of *XRN1*, *DCP2*[2] and *DBP2* [3] deletion on the induction of *GAL1*, *GAL10* and *GAL7* - in part using the strains from Houseley et al. [84]. After overnight growth in glucose(2%) or raffinose(2%), strains were induced with galactose(2%); *GAL1* mRNA was measured over 3 hours of induction and the lag time for *GAL1* expression estimated. *dbp2*Δstrains were found to have significantly reduced lag times compared to WT when induced from glucose, but to be equivalent to WT when induced from raffinose. This lag time difference was lost in *dbp2*Δ / Reb1BSΔ strains. Similar changes in lag time were seen for *dcp2*Δ and *xrn1*Δ strains, while *dcp2*Δ strains also show a change in steady state levels of the *GAL* genes. Reb1BSΔ showed no change in induction dynamics until dense time series (every 10 minutes around induction time) were taken. This revealed statistically significant changes in *GAL10* and *GAL7* lag times between WT and Reb1BSΔ when cells were switched from 2% glucose to 2% galactose. Looking for a mechanism of action, the authors use CHIP to

[1] These are strains in which the H3K4 lysine has been changed to an alanine which cannot be methylated. This is of course a global change that is not confined to the *GAL1-10* locus

[2] Both involved in lncRNA degradation

[3] An RNA helicase involved in both mRNA degradation and transcriptional repression

measure the density of the repressive Cyc8 protein over the promoter and 5' region of the *GAL* genes. They found it was reduced in *dcp2*Δ and *xrn1*Δ strains under repressive conditions, and hypothesise that this is due to displacement of the Cyc8 repressive complex by lncRNA transcription.

The work of Geisler et al. [57] pre dates that of Cloutier et al. and also investigates the role of Dcp2 and Xrn1 in *GAL* gene regulation. Both *dcp2*Δ and *xrn1*Δ are shown to have increased *GAL10*-lncRNA half lives, increased levels of *GAL10*-lncRNA and delayed *GAL*1 activation. H3K18 acetylation (a general mark of transcriptional activity) was measured in the first 30 minutes after activation and found to be reduced over the whole *GAL1-10* locus in *dcp2*Δ : a chromatin state concomitant with delayed activation. The authors also constructed a *GAL10*-lncRNA knock out strain using a more aggressive approach than Houseley et al.: removing *GAL10* entirely and complementing the knock out with a plasmid harbouring a *GAL10* with no Reb1 binding sites. Inducing these strains from raffinose, the authors found that the removal of *GAL10*-lncRNA partially counteracts the deletion of *DCP2*, indicating that the effect of Dcp2p on transcriptional repression is in some part dependent on the *GAL10*-lncRNA. Though they do not discuss it, when *GAL10*-lncRNAΔ cells with no other perturbations were compared to WT in raffinose induction experiments they were indistinguishable.

That the degradation machinery should influence the *GAL10*-lncRNA repressive effect, which has been shown to act in *cis*, is surprising. The authors hypothesise two possible mechanisms for this observation: R loop [4] mediated repression or a cotranscriptional action by Dcp2p and Xrn1p.

In summary, while Houseley et al. saw a repressive effect that was only significant at low concentrations and mixes of sugars, the other papers seem to find

[4] A 3 strand structure formed between double stranded DNA and a complementary RNA

a significant impact of the lncRNAs under even standard laboratory condition of 2% galactose induction. Pinskaya et al. saw a strong repressive effect in 2% galactose induction that was mediated by *SET1* and the *GAL10*-lncRNAs, but their perturbations were global when compared to the localised mutations of Houseley et al., and this may explain the discrepancy. Geisler et al. similarly use global perturbations throughout their paper and this may explain why they, like Pinskaya et al., see a repressive effect from the *GAL10*-lncRNA even under 2% galactose induction. Though it is not discussed, when their only perturbation was to remove the *GAL10*-lncRNA they did not observe a significant difference from WT cells.

Cloutier et al. [34] appears to conflict with both Houseley et al. and Geisler et al. in that they saw an activating effect due to the *GAL10*-lncRNA, and this was only observed when inducing from repressive conditions. Though this is surprising, a number of factors should be taken into account: Cloutier et al. measured *GAL10* and *GAL7* mRNA, in contrast to Houseley et al. and Geisler et al. who based their conclusions largely on *GAL1* mRNA time series; Cloutier et al. only saw an activating effect for Reb1BSΔstrains when making dense time series measurements of glucose to galactose induction over 5 hours, whereas Houseley et al. only published steady state glucose behaviour and time series for 2% galactose induction from raffinose, and that only over 40 minutes. As such, the activating behaviour seen by Cloutier et al. is not incompatible with the data published by Houseley et al..

From our literature review we see that for the *GAL10*-lncRNA both the regime and mechanism of action is contested, but that no results are in direct conflict. This lead us to believe that further understanding could be gained from the high time resolution single cell data and well controlled environment afforded by our microfluidic system. We requested Reb1BSΔ and WT strains with Gal1p-

GFP fusions from the Tollervey lab, intending to first recapitulate the result of Houseley et al. [84] in our microfluidic device.

5.2 Investigating The Effect *GAL10*-lncRNA on the Induction of *GAL1*-GFP by Time Series Microscopy and Microfluidics

Initially we acquired data for Reb1BSΔ and WT strains exposed to the media used in Houseley et al. [84]: YEP with 2% raffinose/ 0.01 % galactose/ 0.02% glucose. A modified version of the ALCATRAS[39] device was used which allowed up to three strains to be imaged while simultaneously exposed to the same media. To match the protocol of Houseley et al. as closely as possible, cells were grown overnight in YEP 2% raffinose, rediluted in fresh media to an OD of 0.05 in the morning and loaded in the device at an OD of approximately 0.2 . Once loaded, the media was changed to YEP with 2% raffinose/ 0.01 % galactose/ 0.02% glucose and this media maintained for the duration of the experiment. Due to the particulars of the modified microfluidic device, it was not possible to observe the arrival of inducing media, and so the exact time at which cells started experiencing galactose is between 0 and 20 minutes before the first time point of the acquisition. A full protocol for the experiment is given in chapter E.

The results from one of the initial experiments are shown in figure 5.4. It can be seen that in this case they reflect those reported by Houseley et al. and also show a significant later difference between the two strains. Unfortunately this result could not be repeated. This was in part due to the technical difficulties. The YEP media used was highly autofluorescent, making early differences in expression difficult to detect, while the late high peak seen in panel A of figure 5.4 was not

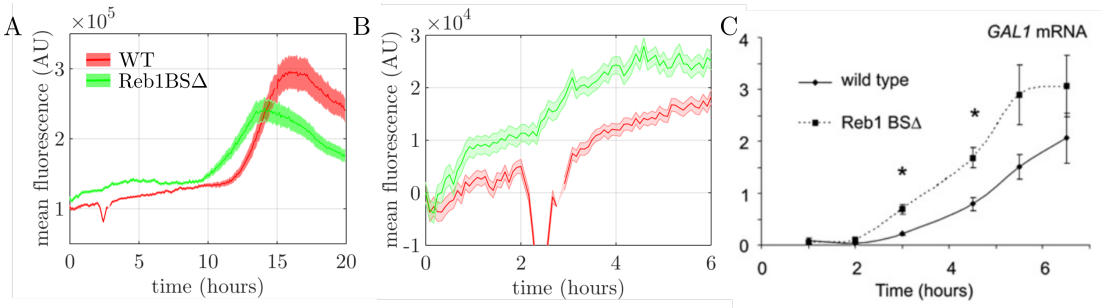

Figure 5.4: Results of WT and Reb1BSΔ cells induced with YEP 2% raffinose/ 0.01 % galactose/ 0.02% glucoseand monitored in the microfluidic device. As described in the main text, cells were loaded into the device in two separate chambers and then exposed to the same inducing media used by Houseley et al.. The exact start time of induction could not be measured but was within 20 minutes of the first time point acquired. Cells were identified, tracked and segmented using the automated algorithms described in chapter 3, with only cells present for more than 220 time points being considered in the analysis. This resulted in 433 WT cells and 488 Reb1BSΔ cells for analysis. Panel A shows the mean behaviour of the each population over the whole twenty hours, with shaded bars showing the standard error on the mean. Panel B shows the first 6 hours with the first time point subtracted: equivalent to that reported in Houseley et al. [84], which is reproduced with permission in panel C. Comparison of B and C shows a very reasonable agreement between the two sets of results (ignoring the large dip due to a loss of focus at around 2 hours). In panel A we can additionally see a late strong induction, that occurs at different times for the two populations.

observed again within the 20 hours of induction. It is possible that if we had been able to observe the cells for longer we would have recovered this behaviour, but experiments lasting more than 20 hours are technically challenging.

Although this result could not be recovered, its appearance led us to conclude that the cells were close to a nutrient regime in which expression differences would be observable in our system. To identify a set of conditions that would be more experimentally tractable, but still reveal differences between the cells, we used flow cytometry to test a range of glucose and galactose concentrations in the less autofluorescent synthetic complete (SC) media. Cells were grown overnight in SC 2% raffinose, rediluted to an OD of 0.05 and grown to an OD of 0.2, at which point they were centrifuged and resuspended in SC 2% raffinose / 0.02% glucose and galactose ranging from 0.01 % to 0.1 %. Fluorescence was measured by flow cytometry 2 hours later. Of the conditions tested, induction with SC 2% raffinose/ 0.04% galactose/ 0.02% glucose was the closest to that used by Houseley et al. which showed robust expression and a significant difference between the strains. This media was selected for further investigation. Results from the flow cytometry are shown in appendix E.

5.3 Application of Alternative Induction Media to Wild Type and Reb1BSΔ Cells

Having identified SC 2% raffinose/ 0.04% galactose/ 0.02% glucose as a condition likely to reveal differences between the strains, we subjected wild type (WT) and Reb1BSΔ cells to induction with this media in the microfluidic device. We used the same protocol as previously described but took advantage of the third chamber of the modified device to simultaneously image control cells expressing no fluorescent protein. The results are shown in figure 5.5 and a clear difference is

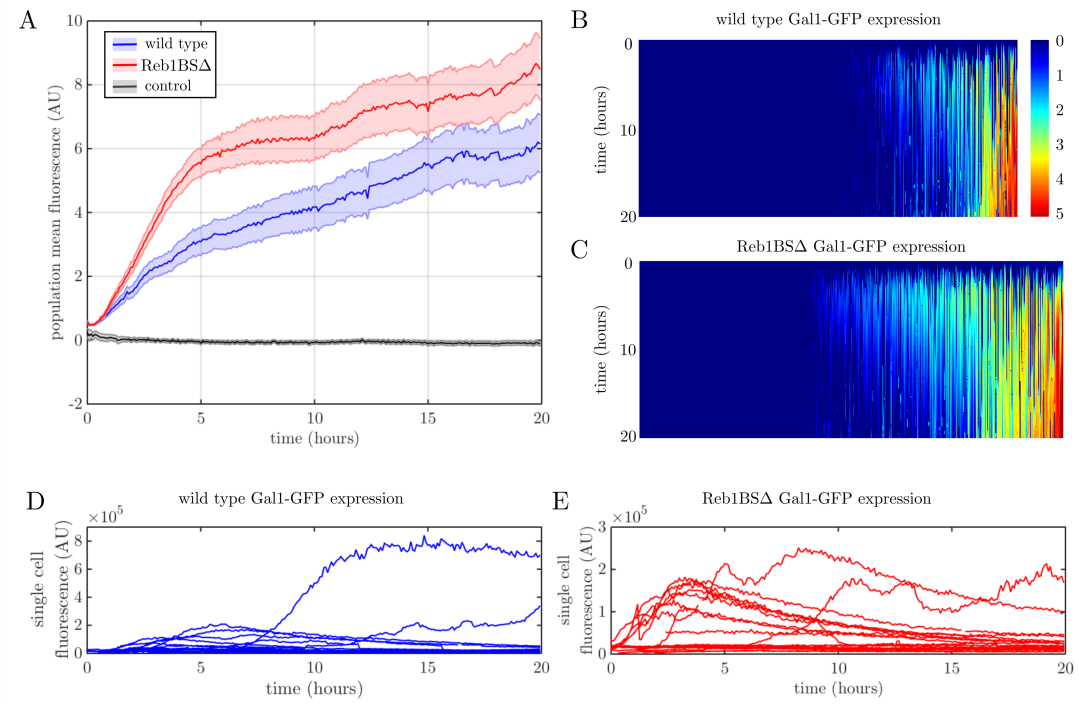

Figure 5.5: Wild type and Reb1Δ induced with SC 2% raffinose/ 0.04% galactose/ 0.02% glucose media show distinct expression patterns. Following the same protocol used previously (see appendix E) wild type, Reb1BSΔ and control cells (expressing no GFP) were exposed to the induction media selected by flow cytometry. Cells were segmented using automated algorithms described in chapter 3 and data corrected for autofluorescence as described in appendix E. 514 wild type, 515 Reb1BSΔ and 28 control cells remained for analysis. Panel A shows the mean behaviour of each strain, with the shaded area showing the standard error on the mean. A clear difference can be seen between the two strains with, as in Houseley et al. [84], the Reb1BSΔ strain inducing earlier and more strongly than the wild type. Panel B and C show kymographs of the logarithm of the fluorescence for single cells. In each case time runs down the y axis and each column is a cell, with the colour at each pixel indicating the log fluorescence of that cell at a given time point. Both kymographs share the same scale given by the bar in panel B. A significant variation in behaviour across each population can be clearly seen. For illustration, a random selection of single cell traces are shown for Wild type (panel D)and Reb1Δ (panel E) strains; colour of the plots have been chosen to match those used in the plots of the population means in panel A. The heterogeneity in both strains can be clearly seen.

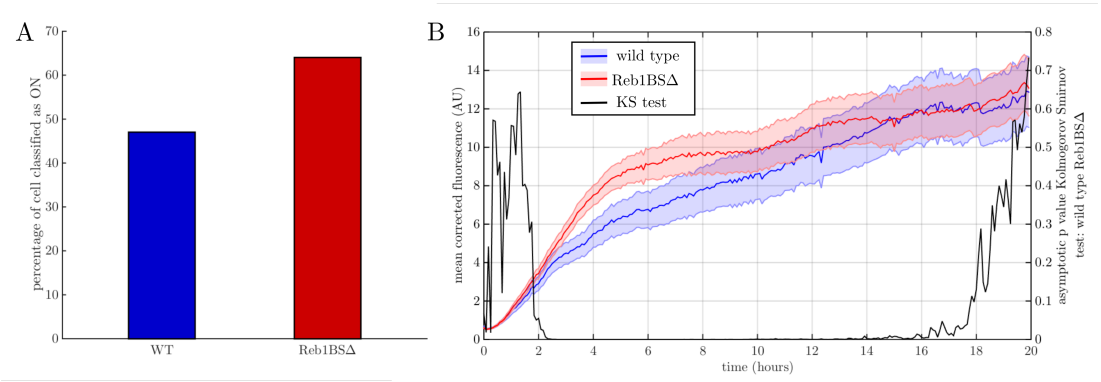

Figure 5.6: Behaviour of wild type and Reb1BSΔcells determined to be ON. Cells from both strains were classified as ON and OFF as described in the main text. Panel A shows the proportion of ON cells in each population. A chi squared test was applied and the proportions found to be statistically significantly different at the 1% significance level. The blue and red plots in panel B shows the mean behaviour for wild type and Reb1BSΔ restricted to just the ON cells; shaded areas show standard errors on the mean as before. The black line shows the asymptotic p value of a time point by time point Kolmogorov-Smirnov (KS) test over the fluorescent data of the ON cells in the two populations. Low p values, as are seen in the middle section of the time series, show that the populations are statistically distinct; high values, as seen at the beginning and end, show that by this criteria they are indistinguishable.

discernible between the average behaviours of the two populations. Looking more closely at data from individual cells, we see that within a population there is also significant variation, and in each case there appear to be two sub-populations: those that induce the *GAL* system and those that do not.

To better understand these two groups we classified cells of both strains as either ON or OFF based on measurements of control cells using a heuristic method. 3 standard deviations of the fluorescence of all control cells across all time points was taken as a threshold. Any cells that crossed this threshold for more than fifty time points (250 minutes) were classified as ON. No control cells were classified as ON by this method, indicating that it is sufficiently stringent that we can be confident of a low false positive rate.

Results from both WT and Reb1BSΔ ON cells are shown in figure 5.6. Different proportions of cells activate the *GAL* network in the two strains, with more cells

switching on in the Reb1BSΔstrain. In panel B we see that even when we restrict our analysis to the ON cells, the two strains still behave differently, but that these differences are temporary and the long term behaviours are indistinguishable.

Since the difference between the ON cells in the two strains appears to be temporary, we reasoned that it may indicate differences in the kinetics of induction in individual cells. To test this hypothesis we calculated a number of kinetic statistics for each trace individually and compared them between the ON cells for wild type and Reb1BSΔ strains. Figure 5.7 depicts the calculation of each statistic. Of these, the KS test identified the lag time and the initial accumulation velocity to be significantly different between the two strains at the 5% significance level.

To summarise the data presented so far: we have acquired data for wild type and Reb1BSΔ strains, both expressing Gal1p-GFP fusions, using our microfluidic device. Looking at the population as a whole, Reb1BSΔstrains show a faster induction of Gal1-GFP and higher expression for the whole 20 hours of observations. Focusing on individual cells we can determine that both populations are heterogeneous and that they can be divided into expressing ON cells and non-expressing OFF cells, with the Reb1BSΔ population having a higher proportion of ON cells. Analysing the two ON subpopulations in more detail we see that even these show different behaviours between the two strains, but that this difference is confined to the early kinetics of induction while the long term expression is the same for the WT and Reb1BSΔ ON cells. This difference in kinetic behaviour was confirmed by statistical analysis of the lag time calculated for each cell, though other statistics such as the time to half maximum fluorescence showed no significant difference.

A faster induction for Reb1BSΔ cells is in accordance with the repressive action

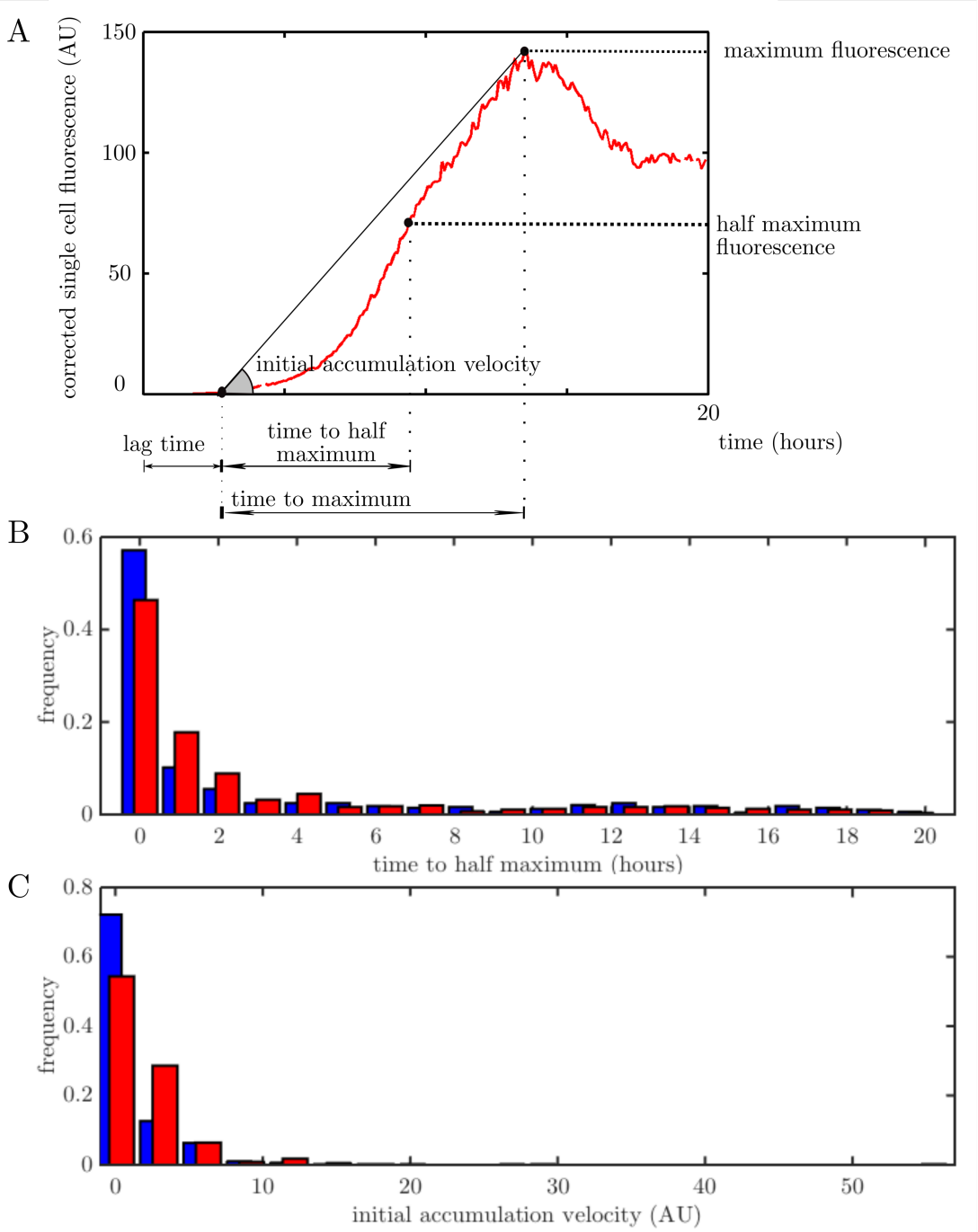

Figure 5.7: Statistics extracted to compare kinetics of wild type and Reb1BSΔ induction.Panel A shows the calculation of each statistic,with each label corresponds to a different quantity that was tested. Those along the x axis are time periods that were compared, those along the y axis are fluorescence values and initial accumulation velocity is an angle calculated as arctan(maximum fluorescence/time to maximum)[34]. Each statistic was calculated for all the ON cells in each population and the KS test applied to determine if the distribution of the parameter was statistically distinguishable. Of the five parameters tested, only the distributions of lag time and initial accumulation velocity were significantly different. These are shown in panels B and C: blue bars show the distribution for wild type cells and red bars that of Reb1BSΔ .

of the *GAL10*-lncRNA observed by Houseley et al.. That the difference is kinetic reflects the result of Cloutier et al. [34], though they saw an inducing effect due to the lncRNA rather than a repressive one. A surprising feature of our data is that the mean fluorescence of the two populations is distinguishable over the entire twenty hour period (figure 5.5). This is not in keeping with Houseley et al. and Cloutier et al., who both saw transient differences between the strains. There are many factors that could be responsible for this, not least the differences in induction media and measurement methodology, but the observation that ON cells also show only transient differences between the two strains implied to us that these discrepancies between our data and the literature might be reconciled. An important distinction between our experiments and those described above is that those studies analysed whole cell culture, which necessarily includes any new born cells. In this sense we observe a different population from the one assayed in cell culture experiments, because we interrogate the same mother cells over the entire twenty hour period.

An advantage of the microfluidic device is that we are also able to observe birth events, and to correlate these with the fluorescence of mothers, which can help us to bridge the gap between our data and that from cell culture. Doing this by hand is prohibitively laborious, so we employed automated daughter identification scripts written by M Crane in the lab. This work is still in development, and the algorithms have not yet been refined or thoroughly characterised, but informal inspection indicates that the data produced is sufficiently accurately for broad conclusions.

Figure 5.8 shows the birth rate data so obtained, which reveal that in both populations a significantly higher birth rate is observed for ON cells than for OFF cells, but that no difference in birth rate can be confirmed between the two strains. This conclusion has two exciting corollaries. First, it shows that by mediating

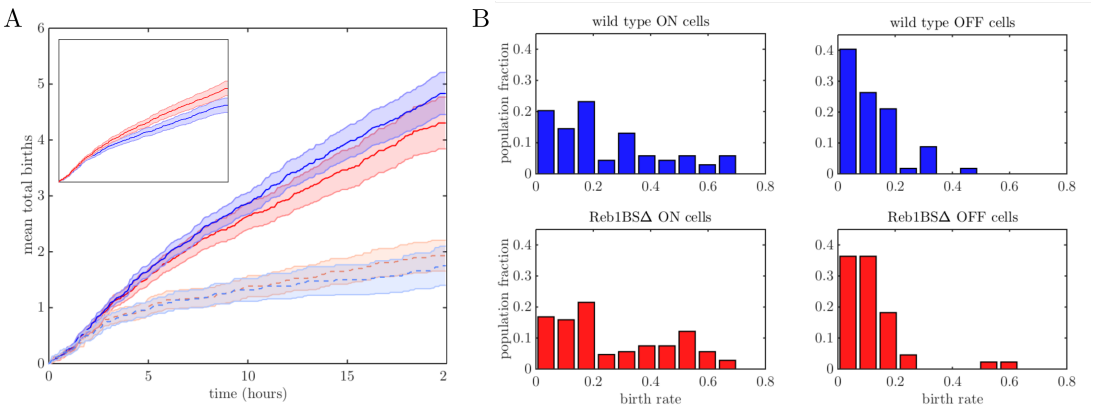

Figure 5.8: Automated scripts were applied to the microfluidic data for wild type and Reb1BSΔ strains to obtain approximate times of birth events for mother cells. Panel A shows the mean total birth events plotted against time. As before, the wild type is plotted in blue while the Reb1BSΔ is plotted in red. The inset is the mean over the population, while the main plot is the population divided into ON cells (upper line in both cases) and OFF cells (lower line). It can be seen that in both strains the birth rate is significantly higher for ON cells than OFF cells. Panel B shows histograms of birth rates (number of daughters/duration present for each mother) for the four sub-populations. KS tests showed a significant difference between ON cells and OFF cells in all cases, but not between the ON cells of wild type and those of Reb1BSΔ or the equivalent sub-populations for OFF cells.

ON/OFF transitions the *GAL10*-lncRNA could impact fitness in different sugar regimes. This, to our knowledge, has not been shown in the literature. Secondly, it may provide a route to aligning our results with those of groups working in batch culture. If ON cells have a higher birth rate, the ON state is inherited (which inspection of the images indicates) and ON cells of the two strains have indistinguishable behaviour at later times then this could explain why the effect of the *GAL10*-lncRNA is transient in batch culture. More importantly, it could have bearing on the population level impact of the lncRNA and the evolutionary advantage it imparts.

Conclusions connecting the lncRNA with fitness require that the activation of the *GAL* system increases birth rate in the induction media. This has not been shown in our experiments, and it could certainly be the case that causality is actually reversed: dividing cells are more prone to activate the *GAL* network whether it

is beneficial or not. To investigate this further, and to understand more about the transcriptional dynamics of *GAL1* we constructed strains in which the Gal1p coding sequence was removed and replaced by the fast folding/fast degrading UBI-MΔk-GFPγ reporter described in chapter 2.

5.4 Investigation of Transcriptional Dynamics Using the Fast Transcriptional Reporter UBI-MΔk-GFPγ

To further investigate both the utilisation of galactose and transcriptional dynamics we constructed two new strains: wild type and Reb1BSΔ cells in which the *GAL1* open reading frame was replaced with the UBI-MΔk-GFPγ transcriptional reporter. For brevity these strains will hereafter be referred to as WT* and Reb1BSΔ*. They were constructed in procedure described in chapter 2: standard lithium acetate transformation of PCR amplified coding sequences targeted to the *GAL1* open reading frame.

Since these strains no longer express Gal1p, they should be unable to metabolise galactose [44] and comparison of their birth rates should show whether the difference in birth rate between ON and OFF cells seen in WT and Reb1BSΔ cells is due to galactose utilisation, or if cells that are dividing are more likely to express the *GAL* genes independently of any benefit. At the same time, the expression of UBI-MΔk-GFPγ from the Gal1 promoter may help us understand the underlying transcriptional dynamics and provide a subject for promoter model inference. Figure 5.9 shows results from the automated budding event analysis performed on data from WT* and Reb1BSΔ* strains. Due to differences in the expression

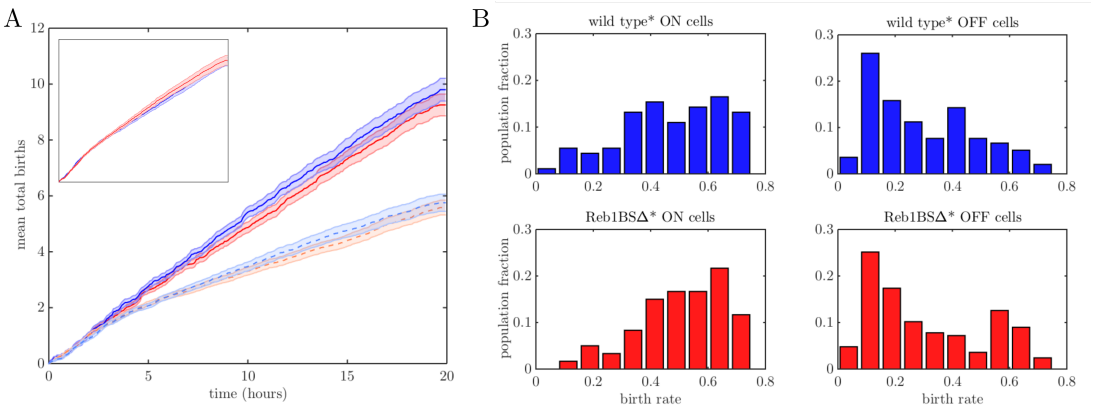

Figure 5.9: Birth statistics for wild type* and Reb1BSΔ* cells. Birth rate data was extracted and processed as in 5.8, with the caveat that the definition of ON and OFF cells was modified to take account of changes in the data (see later discussion). As before, panel A is the mean total birth rate for wild type* (blue) and Reb1BSΔ* (red) strains with cells separated into ON (the higher lines) and OFF (the two lower lines). Despite their inability to metabolise galactose, ON cells in both strains show a higher mean birth rate. Panel B shows the histograms of the birth rates of the two cells separated in ON and OFF cells. As in the previous data set, the ON and OFF cells are significantly different across the strains but there are no detectable differences between strains wild type* and Reb1BSΔ*.

patterns cells were classified as ON or OFF in a slightly different way which will be discussed shortly.

The results show that despite the absence of the Gal1p protein, which is reported to be vital for galactose metabolism [44], the ON and OFF cells still display disparate birth rates similar to those seen in WT and Reb1BSΔ cells. Although confirmation is necessary, this implies that it is the high birth rate of cells which causes *GAL* gene expression rather than the expression of the *GAL* genes that facilitates a high birth rate.

Why the cell birth rates would respond in such a non-uniform way to the induction media when the cells do not appear to be utilising galactose is not clear, and further investigation is certainly required before any conclusions can be drawn, but it seems likely that the difference is a pre-existing feature of the population and it would be interesting to investigate the mechanism for this. The fact that the causality is the opposite of what one might assume does not prevent the

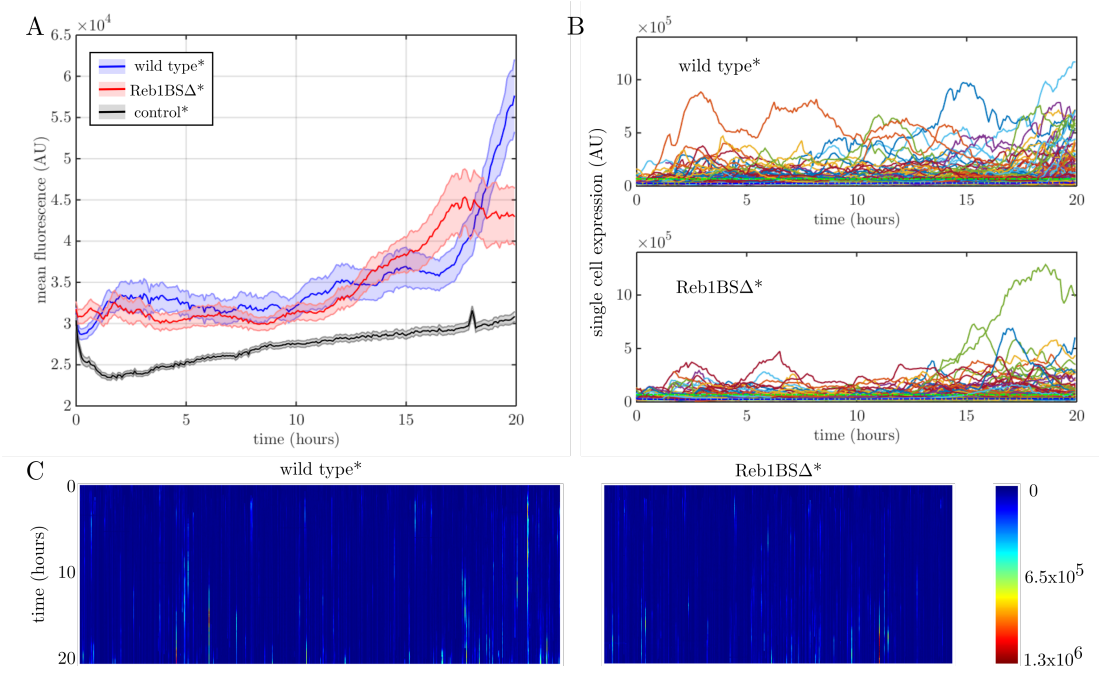

Figure 5.10: Fluorescence data for control*,wild type* and Reb1BSΔ* cells. Panel A shows the mean fluorescence over the whole population, plotted against time, with shaded areas showing standard error on the mean. The behaviour is very different from that of wild type and Reb1BSΔ , with clear induction only at the very end of the experiment. Panel B shows individual traces from each strain and C shows the single cell fluorescence of each population as a kymograph. In contrast to figure 5.5 the raw fluorescence value is plotted rather than the logarithm, which was more informative in this case. Comparison of both panels with their equivalents in figure 5.5 show that the behaviour is quite different. Rather than a classic induction to a steady state the expression appears to be pulsatile, with cells displaying transient periods of expression.

correlation explaining why we observe lasting differences while other studies see transient ones. If dividing cells are more likely to express *GAL*, the requisites will still be fulfilled for ON cells coming to dominate the population. What does seem unlikely from this data is fitness effect due to the lncRNA in these conditions.

We now turn to the fluorescence data, which is shown in figure 5.10. Processing and correction of data differ from those applied to the wild type and Reb1BSΔ and are explained in appendix E. A different control strain was also used, in which the *GAL1* ORF was replaced with the UBI-MΔk-mKate2 coding sequence. This

is referred to as control*.

The data is clearly different from that shown in figure 5.5, with fluorescence throughout the experiment but no clear point of induction. Inspection of single cells shows that the fluorescence is pulsatile, with transient periods of expression rather than clear induction. As with the *gal1*Δ strains (wild type and Reb1BSΔ) a number of statistics were developed to try and distinguish the two populations but none were found to be significantly different between the two strains.

It should be noted that in these experiments we are using a different reporter (free UBI-MΔk-GFPγ as oppose to EGFP tagged to the Gal1 protein), and this alone could lead to a different appearance for the traces. UBI-MΔk-GFPγ is engineered to have a shorter half-life, and this would lead to a lower mean expression and higher coefficient of variation [139], which might give the traces a more pulsatile appearance. However, as shown in chapter 2 the half life of UBI-MΔk-GFPγ is only about half that of an unmodified EGFP, and is therefore unlikely to explain the factor of 10 change in the fluorescence of the cells. In the absence of any sort of feedback the decay rate would also make little difference to the shape of the mean fluorescence of the population over time, as it acts largely as a scaling factor. Taken together these two observations show that though the change of reporter is important, it cannot alone explain the difference in the results between the two experiments, and there must therefore be some change in expression of the gene.

This interesting behaviour has not, to our knowledge, been previously reported. Kaufmann et al. [96] have observed slow stochastic switching of the *GAL* network, but this required the Gal80 repressor to be expressed from a constitutive promoter, abrogating the negative feedback loop in which it is involved. Comparing these results with those presented in figure 2.15 of chapter 2, we see that for similar *gal1*Δ cells in SC 2% galactose the induction is more sustained, leading us to believe that this effect must be a combination of the particular induction

media and the Gal1 deletion. As we have discussed, at high concentration Gal1p can replace Gal3p as the dominant transcriptional co-activator, and the difference between the *GAL1* (wild type and Reb1BSΔ) and *gal1*Δ (wild type* and Reb1BSΔ*) strains could be explained by the impairment of this positive feedback loop. This might also explain why *GAL1* cells are easily distinguishable while *gal1*Δ cells are not. If the Gal1p positive feedback is responsible for the sustained inductions in *GAL1* cells, then a small change in its expression due to the *GAL10*-lncRNA would be amplified; without this amplification, the difference between wild type* and Reb1BSΔ* cells may be harder to see.

To try and identify changes in *GAL1* transcription that could confirm this hypothesis, we applied a model based Bayesian inference scheme to the data from *gal1*Δ cells presented in this section. We reasoned that if we could infer a model of transcription from the Gal1 promoter, with and without transcription of the *GAL10*-lncRNA, we could test its behaviour in a simple positive feedback loop and challenge this hypothesis, that Gal1p mediated feedback was responsible for the differences in expression kinetics.

5.4.1 Application of DPP Inference Scheme to *gal1*ΔStrains

Various of inference algorithms exist, employing different approximations and sampling techniques [51, 130, 193, 195]. The DPP algorithm of Zechner et al. [195] is an efficient algorithm that makes no approximations, can infer extrinsically varying parameters and has a convenient matlab implementation. For these reasons it was selected for a first attempt at inference. In preparing our data we followed the protocol of the original paper as closely as possible, the details of which are given in appendix E. We also curated by hand the tracking and

segmentation of approximately 80 cells from each strain to minimise errors.

We quickly found that the algorithm was too slow to infer parameters even for the subset of 80 cells, and so we reduced further to a set of 10 cells from each strain in order to be able to run the inference algorithm within a day. To try and preserve the population structure as much as possible, the 80 original cells were subdivided into ON and OFF cells and cells sampled from separate groups in proportion to the size of the ON/OFF sub-populations in the whole population. This was done to preserve the population ratio of ON to OFF cells in the sub sample to which inference was applied. Inference was run on each of the strains separately with the intention of comparing the parameters. A simple two state model was used as in Zechner et al. [195].

In addition to the inference on the real data, we wanted to test the performance of the algorithm on simulated data subject to measurement errors similar to those we expect in our system. Drawing on investigation of error sources in earlier chapters, we generated a data set of 10 cells as follows. The master equation for a two state model was simulated using the stochkit2 stochastic simulation package [151] (all parameters available in appendix E). This raw data was multiplied by a scaling factor to give fluorescence from protein, and then multiplied by a separate scale factor for each cell to represent size effects. To this we added log normal noise and then Poissonian, and these fluorescent values were then added to the traces from control* cells to simulate autofluorescence. The measurements were then treated for inference in the same way as the experimental data, giving the fluorescence-protein scaling factor 20 % error to reflect imperfect estimation of this parameter. The DPP algorithm was applied to both data sets (details in appendix E), the results for the intrinsic kinetic parameters and measurement noise are shown in figure 5.11.

From the results on the simulated data set it could be said that the performance is reasonable, the true values usually falling within an order of magnitude of the

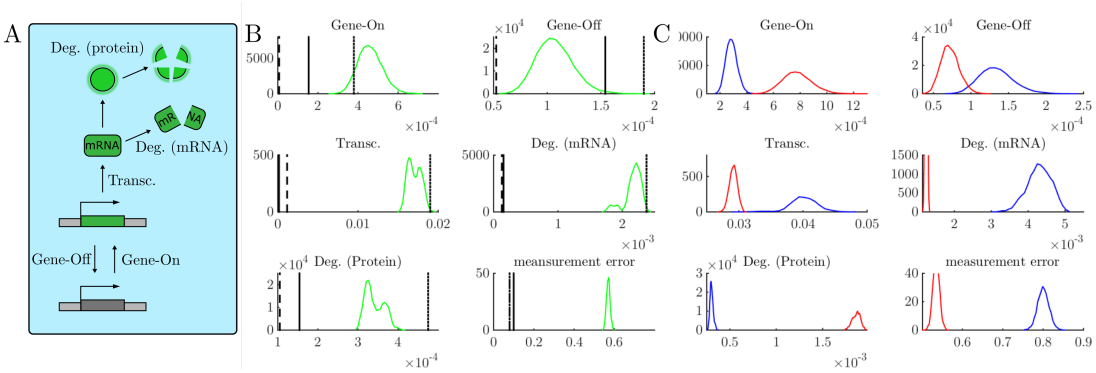

Figure 5.11: The DPP [195] algorithm was applied to data from wild type* and Reb1BSΔ*, as well as simulations with reasonable microscope errors. Panel A is a depiction of the two state model used in the inference and conveys the meaning of each parameter. Translation rate is not shown since this is fitted as an extrinsic variable and was not considered in the analysis. Panel B shows the results of simulated data for 10 cells and 240 time points at 5 minute intervals, equivalent to our experimental data. A range of errors were added which is detailed in the main text. The green line is the inferred posterior for each parameter, the solid black line is the true value used in simulation, while the dashed black lines mark the 5th and 95th percentile of the prior distribution. It can be seen that though the algorithm performs reasonably well, the true value often falls outside the high a posteriori region. Panel C shows the result of applying the inference algorithm to wild type* (blue line) and Reb1BSΔ*(red line) data. 10 cells and all 240 time points were used in each case.

maximum a posteriori value. However, it is clear that the high a posterior region does not always include the true value, and it is therefore difficult to compare the inferred parameters for wild type* and Reb1BSΔ* cells in a meaningful way. This might be solved by more computational time or it could be a short coming of the data, and more simulations would be required to understand which of these two is the cause of the poor performance. However, each of these data sets required a whole night to run on a desktop computer, so exploring our experimental regime by simulation is infeasible. This slow performance is most likely due to the long duration of the experiments and the large number of proteins, which can slow down the Gillespie algorithm considerably. If we intend to continue with inference, our large data sets and long time series will require algorithms that make use of some of the many approximations to the master equation that are commonly used [193]. Adapting inference techniques to manage our large and extrinsically variable datasets present a technically challenging problem, that is unlikely to be solved by established algorithms.

5.5 Discussion

The work presented in this chapter is a preliminary application of the transcriptional reporters, error analysis, microfluidic device and associated segmentation software. Using these tools we were able to observe subtle, kinetic differences in induction between wild type and Reb1BSΔ strains, and to associate these differences with single cell birth rates and fitness. Our application of the transcriptional reporter UBI-MΔk-GFPγ is still in the early stages, particularly with regards to model inference algorithms, but the observation of distinct behaviour from the *GAL1* strains implies that these strains could reveal a great deal about regulation of the *GAL* system in these media conditions.

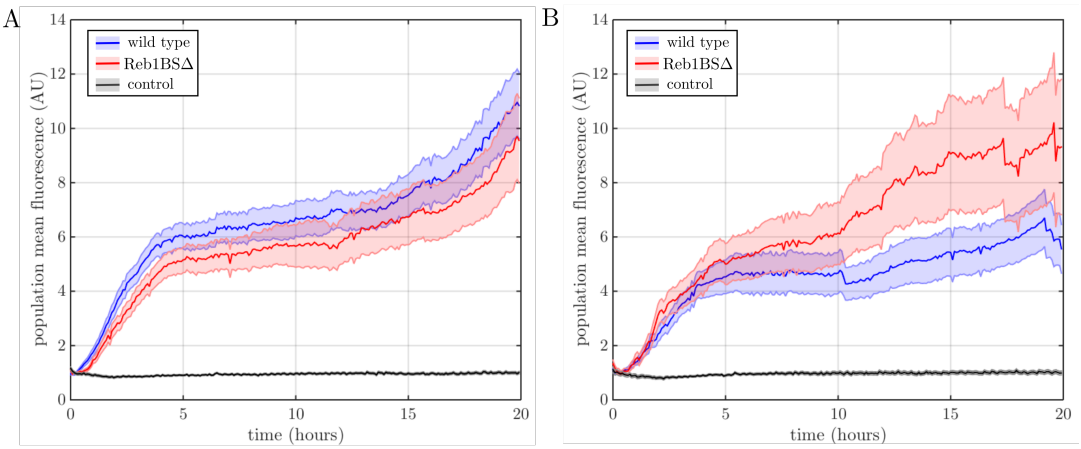

Figure 5.12: Attempted repeats of galactose induction experiments of wild type and Reb1BSΔ strains. As in figure 5.5, strains were grown overnight in SC raffinose (2%) and then induced with SC 2% raffinose/ 0.04% galactose/ 0.02% glucose media. Panel A and B show two attempts at repeats done on different days, with 900 and 485 cells respectively. Clearly these are significantly different from the result shown in figure 5.5, and we currently believe that this is due to both variation in overnight culturing and the sensitivity of the result to the number of cells that can be observed. Both these issues are being addressed currently in the lab.

The priority now is to acquire dependable repeats of the experiments discussed here, and controls to confirm the dependence on the *GAL10*-lncRNA. While we have pursued this vigorously, technical difficulties have so far prevented us from obtaining a set of convincing replicates. Once this has been overcome, there are a number of experiments that could advance the project and understanding of the *GAL* system. The connection between the dividing and expressing subpopulations would be of great interest to pursue further. Particularly, whether the cells that continue dividing in the new media show higher growth in pure raffinose media before the switch. This experiment could be performed using the dynamic media control afforded by a recently developed modified microfluidic protocol, which would also allow us to later remove the galactose and monitor the response of the cells.

Looking at the difference between the *GAL1* and *gal1*Δ cells, it would be revealing to apply the same analysis to heterozygous diploids with one UBI-MΔk-

GFPγ reporter and one functional Gal1 gene. Heterozygous diploid production is a straightforward protocol, and a modified diploid device is already in use in the lab. This would provide a means to test the positive feedback hypothesis outlined in this chapter.

Further application of inference to the *gal1*Δ strains will require significant additional work, and it is likely that robust inference over the large data sets we generate will prove to be a challenging inference problem rather than a simple application of existing technology. That said, there is significant expertise on this subject in Edinburgh and in the interdisciplinary SynthSys centre itself. Achieving a full pipeline of large scale quantitative data acquisition and robust statistical inference will be challenging but it is certainly achievable, and the data presented in this chapter shows that the insights obtained could be wide ranging. Even confining ourselves to the *GAL1* gene we see that there are many regulatory mechanisms, with wide ranging relevance [17, 177], that could be straightforwardly perturbed and inspected using the microfluidic system.

In conclusions, the results from this study on the *GAL10*-lncRNA is a good indication that the combination of our well developed microfluidic system and the genetically tractable and well understood *Saccharomyces cerevisiae* model organism can deliver novel and widely relevant results. Though we have only looked at a small portion of the *GAL* regulon here, it seems likely that a similar investigation of the many other model systems in yeast could also produce interesting findings.

Chapter 6

Conclusion

In this thesis we have described work advancing both the utility and understanding of the ALCATRAS microfluidic system. Further, we have shown preliminary data and analysis from its application to the regulation of the *GAL1* gene in *Saccharomyces cerevisiae* , displaying some the insights it can provide. This system, and others like it are, becoming increasingly popular and much of the analysis we have developed is widely applicable.

In chapter 2 we undertook a detailed study of sources of error in our experimental system and the corrections that could be made to mitigate them. A novel protocol was developed which leveraged the microfluidic system to provide a better estimate of flat field correction, and errors arising due to both the camera and the optics were assessed by a combination of measurement and simulation. This revealed that the commonly assumed Poissonian noise is not appropriate for all camera settings, and that this relatively simple source of error may not be as significant as the slow varying systematic errors arising from the microscope optics and the distribution of fluorescence throughout the cell. Based on these

quantitative error estimates, a number of imaging conditions were assessed and compared, with the optimum found to depend on the experimental conditions and protein expression expected. This work illustrates the numerous sources of error which, if not properly considered, could significantly impact on the conclusions of quantitative analysis. It is important to now confirm this computational work with experimental data. Once this is done, we can begin to construct well informed corrections and devise optimal statistics that will increase the accuracy of both our microscope measurements and error estimates.

In the second half of chapter 2 we developed two new fluorophores designed for monitoring transcriptional dynamics. These reporters were constructed and their brightness, degradation rate and maturation rates measured in the most physiological conditions possible. Though the properties of the UBI-MΔkmKate2 reporter were disappointing, the UBI-MΔkGFPγ was a significant improvement on the UBI-MΔkGFP* reporter from which it was derived. Having developed this tool we would now like to apply it to understanding transcriptional dynamics, as we have done in chapter 5. further characterisation of the Fluorophore, especially improving the estimate of its maturation rate, will be important in these applications. We believe that the higher time resolution afforded by our microfluidic device will facilitate this.

The original intention of this work was to develop a pair of distinguishable transcriptional reporters, and this would still be desirable. One possible avenue would be to focus on the CFP fluorophore, that has been shown to respond to the UBI-MΔk tag [69], but since it is not easily distinguishable from GFP it is likely a partner reporter would also have to be developed. Using high throughput DNA fabrication techniques, accessible through the newly established Edinburgh genome foundry, would increase the chance of finding a successful pair of reporters. These would allow intrinsic and extrinsic noise in promoter dynamics to be investigated, which would be of great value in understanding the biological

source of the putative promoter states reported in the literature [74, 172].

In chapter 3 we developed an active contour algorithm to improve the automated segmentation of the images produced by our microscope. Further, we developed a suite of analysis metrics by which to automatically test and compare the performance of segmentation and tracking algorithms using a curated ground truth. Image segmentation is a difficult problem and the field is fragmented, at least in its application to biology. The development of robust tools to easily compare algorithms could be of great value in guiding both users and developers. The metrics described also distinguish different types of error, allowing the user to make informed decisions based on their experimental priorities. Applying these characterisation methods to the different segmentation algorithms developed in the lab, we showed that the active contour algorithm improves the segmentation result and significantly reduces measurement error.

Automated software can always be improved, and as such there is of course more work that could be done on the segmentation software, but a number of straightforward avenues for improvement present themselves. The classifier used to identify cell centres currently only uses features based on the in focus DIC image, and makes no use of the out of focus images that are now routinely acquired in our experiments. These appear to be more informative than the in focus image alone, and rewriting the software to use these images could result in a substantial improvement in cell identification. The cost function used in the active contour algorithm currently imposes shape information only as two heuristically motivated derivatives, but we now have a large collection of curated results that can be used to construct a data driven shape model. This could improve segmentation, and in particular make the result more robust in cases where image quality is poor.

In chapter 4 we developed tools to infer protein-fluorescence ratio and measurement error from stochastic variation in cellular photobleaching. For the error free

case, deterministic estimates of the parameters were derived, and for measurements with errors an approximate likelihood was found and an efficient method for computing it implemented. In application to real data the model of a simple first order photobleaching rate was found to be generally inappropriate. Experimental conditions were sought in which photobleaching was well approximated by a simple first order reaction, and this necessitated developing new tools to quantitatively account for autofluorescence. Though the performance of our estimator was not satisfactory, pursuing this project has increased our understanding of our experimental system and many of the correction methods developed have improved the accuracy of our microscopy data generally. In the project up to this point we have attempted to fit our experimental conditions to our model, and in the future it may be fruitful to instead try and adapt our inference scheme to the bi-exponential process that we generally observe. Though this will probably require a more sophisticated inference methodology than the one we have applied, it is still a simple linear model and would certainly be amenable to inference.

In chapter 5 we applied the experimental system developed to understanding the role of the *GAL10*-lncRNA in Gal1 regulation. We were able to corroborate a repressive effect due to the *GAL10*-lncRNA and reveal its importance in a heterogeneous response that had not been previously observed. Applying analysis methods developed by M Crane, we were able to correlate induction of the *GAL* cluster with birth rate at the single cell level, and thereby relate our microfluidic results to those in batch culture. We constructed new strains in which the *GAL1* open reading frame was replaced with the UBI-MΔkGFPγ reporters developed in chapter 2. Since *gal1*Δ cells are auxotrophic for galactose [44], this allowed us to investigate both the underlying transcriptional dynamics of the *GAL1* gene and the metabolic benefit of galactose utilisation. Comparing the single cell birthrates of the *gal1*$\Delta$and *GAL1* strains we saw no statistical difference, indicating that the correlation between birthrate and *GAL* expression we observed was not caused by

the advantageous utilisation of galactose in a subset of cells, but rather that some other property of those rapidly dividing cells allows for the induction of the *GAL* system. This is an interesting and unexpected result that could challenge the view of *GAL* induction as a calculated response to maximise the energy available to the cell.

Looking at the fluorescence data from the *gal1*Δ strains we saw a pulsatile behaviour different from the clear stable induction seen in the *GAL1* cells, which we hypothesise may be due to the abrogation of the positive feedback loop caused by the regulatory role of Gal1p. Preliminary attempts to apply the dynamic prior propagation (DPP) Bayesian inference algorithm of Zechner et al. [195] indicate that simulation based methods will most likely be too slow for data sets of our size, and algorithms employing approximations of the master equation will be necessary.

Confirming these results by reliable repeats and thorough controls are necessary, as is refinement and characterisation of the daughter counting algorithm employed, but the data presented already shows that the microfluidic system is capable of elucidating subtle phenomena in gene regulation and relating expression to fitness. Looking forward, it would be interesting to further investigate the connection between birthrate and gene expression, particularly in a greater range of galactose concentrations, and to see if those cells dividing slowly in the presence of galactose show a slower rate of division in the raffinose media to which they are initially exposed. This was not possible in the experiments presented here but could be done using alternative microfluidic devices available in the lab. Further work on the *GAL10*-lncRNA would also be informative, particularly investigating some of the other roles ascribed to it in the literature and looking at its effect on expression in dynamic environments. Chromatin state has been implicated in short term reinduction [99] memory, and it seems quite possible that the *GAL10*-lncRNA, with its influence on promoter chromatin, could be

important in this process.

Turning to the application of Bayesian inference algorithms to our data, employing approximation of the master equation would be most conducive to progress. Though performing timely and rigorous inference on our large datasets is likely to prove challenging, the benefits of combining high quality single cell data acquisition and robust model inference are potentially enormous.

The microfluidic device employed in the lab is a sophisticated system, and there is still a great deal of work to be done to refine and characterise the experimental protocol and associated analysis, but as we have already demonstrated the potential insights are manifold. Biology is inexorably becoming a quantitative science of complex systems, and in this environment high quality well characterised data, that can be used not only to infer models but also to estimate the confidence one can attach to them, will be invaluable. Already people are attempting to integrate results from different areas of biology into complete descriptions of organisms [94]. This must be seen as the eventual aim of molecular and cellular biology, but it will only be possible in a rigorous way if the component models are well founded and their limitations are understood. The experimental and analytical methodologies developed here, combined with the genetic tractability and experimental facility of yeast, has the potential to make significant contributions to this challenging endeavour.

Bibliography

[1] Dariusz Abramczyk, Stacey Holden, Christopher J. Page, and Richard J. Reece. Interplay of a ligand sensor and an enzyme in controlling expression of the Saccharomyces cerevisiae GAL genes. *Eukaryotic Cell*, 11(3):334–342, 2012.

[2] Murat Acar, Attila Becskei, and Alexander van Oudenaarden. Enhancement of cellular memory by reducing stochastic transitions. *Nature*, 435 (7039):228–32, May 2005.

[3] Raluca Apostu and Michael C Mackey. Mathematical model of GAL regulon dynamics in Saccharomyces cerevisiae. *Journal of theoretical biology*, 293: 219–35, January 2012.

[4] a Bachmair and a Varshavsky. The degradation signal in a short-lived protein. *Cell*, 56(6):1019–1032, 1989.

[5] Monya Baker. RNA imaging in situ. *Nature Methods*, 9(8):787–790, July 2012.

[6] Nathalie Q Balaban, Jack Merrin, Remy Chait, Lukasz Kowalik, and Stanislas Leibler. Bacterial persistence as a phenotypic switch. *Science (New York, N.Y.)*, 305(5690):1622–5, September 2004.

[7] Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin

O'Shea, Yitzhak Pilpel, and Naama Barkai. Noise in protein expression scales with natural protein abundance. *Nature genetics*, 38(6):636–43, June 2006.

[8] Attila Becskei, Benjamin B Kaufmann, and Alexander van Oudenaarden. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature genetics*, 37(9):937–44, September 2005.

[9] Matthew R Bennett and Jeff Hasty. Microfluidic devices for measuring gene network dynamics in single cells. *Nature Reviews Genetics*, 10(9):628–638, September 2009.

[10] Matthew R Bennett, Wyming Lee Pang, Natalie a Ostroff, Bridget L Baumgartner, Sujata Nayak, Lev S Tsimring, and Jeff Hasty. Metabolic gene regulation in a dynamically changing environment. *Nature*, 454(7208):1119–1122, August 2008.

[11] Mojca Benčina. Illumination of the spatial order of intracellular pH by genetically encoded pH-sensitive sensors. *Sensors (Basel, Switzerland)*, 13 (12):16736–58, January 2013.

[12] Andrew J Berglund. Nonexponential statistics of fluorescence photobleaching. *The Journal of chemical physics*, 121(7):2899–903, August 2004.

[13] Sara Berthoumieux, Hidde de Jong, Guillaume Baptist, Corinne Pinel, Caroline Ranquet, Delphine Ropers, and Johannes Geiselmann. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Molecular systems biology*, 9(634):634, January 2013.

[14] N Billinton and a W Knight. Seeing the wood through the trees: a review of techniques for distinguishing green fluorescent protein from endogenous autofluorescence. *Analytical biochemistry*, 291(2):175–197, 2001.

[15] B Birge. PSOt-a particle swarm optimization toolbox for use with Matlab. *IEEE Swarm Intelligence Symposium Proceedings*, 2003.

[16] Andrew Blake and Michael Isard. *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer Science & Business Media, 2012.

[17] David Botstein and Gerald R Fink. Yeast: an experimental organism for 21st Century biology. *Genetics*, 189(3):695–704, November 2011.

[18] Susanne Brandes, Zeinab Mokhtari, Fabian Essig, Kerstin Hünniger, Oliver Kurzai, and Marc Thilo Figge. Automated segmentation and tracking of non-rigid objects in time-lapse microscopy videos of polymorphonuclear neutrophils. *Medical image analysis*, November 2014.

[19] Kristian Bredies and Heimo Wolinski. An active-contour based algorithm for the automated segmentation of dense yeast populations on transmission microscopy images. *Computing and Visualization in Science*, April 2012.

[20] Christopher R Brown and Hinrich Boeger. Nucleosomal promoter variation generates gene expression noise. *Proceedings of the National Academy of Sciences of the United States of America*, 2014, December 2014.

[21] Christopher R. Brown, Changhui Mao, Elena Falkovskaia, Melissa S. Jurica, and Hinrich Boeger. Linking Stochastic Fluctuations in Chromatin Structure and Gene Expression. *PLoS Biology*, 11(8):e1001621, August 2013.

[22] Gene O Bryant, Vidya Prabhu, Monique Floer, Xin Wang, Dan Spagna, David Schreiber, and Mark Ptashne. Activator control of nucleosome occupancy in activation and repression of transcription. *PLoS biology*, 6(12): 2928–2939, 2008.

[23] Ghislain G Cabal, Auguste Genovesio, Susana Rodriguez-Navarro, Christophe Zimmer, Olivier Gadal, Annick Lesne, Henri Buc, Frank Feuerbach-Fournier, Jean-Christophe Olivo-Marin, Eduard C Hurt, and Ulf Nehrbass. SAGA interacting factors confine sub-diffusion of transcribed genes to the nuclear envelope. *Nature*, 441(7094):770–773, 2006.

[24] Andrew P Capaldi, Tommy Kaplan, Ying Liu, Naomi Habib, Aviv Regev, Nir Friedman, and Erin K O'Shea. Structure and function of a transcriptional network activated by the MAPK Hog1. *Nature genetics*, 40(11): 1300–6, November 2008.

[25] Lucas B. Carey, David van Dijk, Peter M. A. Sloot, Jaap A. Kaandorp, and Eran Segal. Promoter sequence determines the relationship between expression level and noise. *PLoS biology*, 11(4):e1001528, January 2013.

[26] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David a Guertin, Joo Han Chang, Robert a Lindquist, Jason Moffat, Polina Golland, and David M Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, January 2006.

[27] J. Chalfoun, M. Majurski, K. Bhadriraju, S. Lund, P. Bajcsy, and M. Brady. Background intensity correction for terabyte-sized time-lapse images. *Journal of Microscopy*, 257(November):226–238, 2015.

[28] Ching-Wei Chang, Dhruv Sud, and Mary-Ann Mycek. Fluorescence lifetime imaging microscopy. *Methods in cell biology*, 81(06):495–524, January 2007.

[29] Rohit Chatterjee, Mayukh Ghosh, Ananda S Chowdhury, and Nilanjan Ray. Cell tracking in microscopic video using matching and linking of bipartite graphs. *Computer methods and programs in biomedicine*, pages 1–10, August 2013.

[30] Long Chen, Leanne Lai Chan, Zhongying Zhao, and Hong Yan. A novel cell nuclei segmentation method for 3D C. elegans embryonic time-lapse images. *BMC bioinformatics*, 14(1):328, November 2013.

[31] Shasha Chong, Chongyi Chen, Hao Ge, and X. Sunney Xie. Mechanism of transcriptional bursting in bacteria. *Cell*, 158(2):314–326, 2014.

[32] Yolanda T. Chong, Judice L.Y. Koh, Helena Friesen, Kaluarachchi Duffy, Michael J. Cox, Alan Moses, Jason Moffat, Charles Boone, and Brenda J. Andrews. Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell*, 161(6):1413–1424, 2015.

[33] Clonetech. Clonetech In-Fusion Cloning System.

[34] Sara C Cloutier, Siwen Wang, Wai Kit Ma, Christopher J Petell, and Elizabeth J Tran. Long noncoding RNAs promote transcriptional poising of inducible genes. *PLoS biology*, 11(11):e1001715, November 2013.

[35] Alejandro Colman-Lerner, Andrew Gordon, Eduard Serra, Tina Chin, Orna Resnekov, Drew Endy, C Gustavo Pesce, and Roger Brent. Regulated cell-to-cell variation in a cell-fate decision system. *Nature*, 437(7059):699–706, September 2005.

[36] Scott Cookson, Natalie Ostroff, Wyming Lee Pang, Dmitri Volfson, and Jeff Hasty. Monitoring dynamics of single-cell gene expression over multiple cell cycles. *Molecular systems biology*, 1:2005.0024, January 2005.

[37] B P Cormack, R H Valdivia, and S Falkow. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, 173(1 Spec No):33–38, 1996.

[38] Adam D Coster, Chonlarat Wichaidit, Satwik Rajaram, Steven J Altschuler, and Lani F Wu. A simple image correction method for high-throughput microscopy. *Nature Methods*, 11(6):602–602, May 2014.

[39] Matthew M. Crane, Ivan B N Clark, Elco Bakker, Stewart Smith, and Peter S. Swain. A microfluidic system for studying ageing and dynamic single-cell responses in budding yeast. *PLoS ONE*, 9(6):1–10, 2014.

[40] Richard N. Day and Fred Schaufele. Fluorescent protein tools for studying protein dynamics in living cells: a review. *Journal of Biomedical Optics*, 13(3):031202, 2008.

[41] Alberto Diaspro, Giuseppe Chirico, Cesare Usai, Paola Ramoino, and Jurek Dobrucki. Photobleaching. In *Handbook of Biological Confocal Microscopy, third edition*, chapter 39, pages 690–702. 2006.

[42] Sotiris Dimopoulos, Christian E Mayer, Fabian Rudolf, and Joerg Stelling. Accurate cell segmentation in microscopy images using membrane patterns. *Bioinformatics (Oxford, England)*, pages 1–8, 2014.

[43] Andreas Doncic, Umut Eser, Oguzhan Atay, and Jan M Skotheim. An algorithm to automate yeast segmentation and tracking. *PloS one*, 8(3): e57970, January 2013.

[44] H C Douglas and D C Hawthorne. Enzymatic Expression and Genetic Linkage of Genes Controlling Galactose Utilization in Saccharomyces. *Genetics*, 49:837–844, 1964.

[45] Kevin W Eliceiri, Michael R Berthold, Ilya G Goldberg, Luis Ibáñez, B S Manjunath, Maryann E Martone, Robert F Murphy, Hanchuan Peng, Anne L Plant, Badrinath Roysam, Nico Stuurmann, Jason R Swedlow, Pavel Tomancak, and Anne E Carpenter. Biological imaging software tools. *Nature Methods*, 9(7):697–710, June 2012.

[46] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science (New York, N.Y.)*, 297 (5584):1183–1186, August 2002.

[47] S. O. Enfors, M. Jahic, A. Rozkov, B. Xu, M. Hecker, B. Jürgen, E. Krüger, T. Schweder, G. Hamer, D. O'Beirne, N. Noisommit-Rizzi, M. Reuss, L. Boone, C. Hewitt, C. McFarlane, A. Nienow, T. Kovacs, C. Trägå rdh, L. Fuchs, J. Revstedt, P. C. Friberg, B. Hjertager, G. Blomsten, H. Skogman, S. Hjort, F. Hoeks, H. Y. Lin, P. Neubauer, R. Van der Lans, K. Luyben, P. Vrabel, and ÅManelius. Physiological responses to mixing in large scale bioreactors. *Journal of Biotechnology*, 85(2):175–185, 2001.

[48] Renan Escalante-Chong, Yonatan Savir, Sean M Carroll, John B Ingraham, Jue Wang, Christopher J Marx, and Michael Springer. Galactose metabolic genes in yeast respond to a ratio of galactose and glucose. *Proceedings of the National Academy of Sciences of the United States of America*, January 2015.

[49] Frederik Faden, Stefan Mielke, Dieter Lange, and Nico Dissmeyer. Generic tools for conditionally altering protein abundance and phenotypes on demand. *Biological Chemistry*, 395(7-8):737–762, 2014.

[50] D Falconnet, a Niemistö, R J Taylor, M Ricicova, T Galitski, I Shmulevich, and C L Hansen. High-throughput tracking of single yeast cells in a microfluidic imaging matrix. *Lab on a chip*, 11(3):466–473, 2011.

[51] Bärbel Finkenstädt, Dan J. Woodcock, Michal Komorowski, Claire V. Harper, Julian R E Davis, Mike R H White, and David a. Rand. Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: An application to single cell data. *Annals of Applied Statistics*, 7(4):1960–1982, 2013.

[52] Monique Floer, Xin Wang, Vidya Prabhu, Georgina Berrozpe, Santosh Narayan, Dan Spagna, David Alvarez, Jude Kendall, Alexander Krasnitz, Asya Stepansky, James Hicks, Gene O. Bryant, and Mark Ptashne. A

RSC/nucleosome complex determines chromatin architecture and facilitates activator binding. *Cell*, 141(3):407–418, 2010.

[53] Erwin Füreder-Kitzmüller, Jan Hesse, Andreas Ebner, Hermann J. Gruber, and Gerhard J. Schütz. Non-exponential bleaching of single bioconjugated Cy5 molecules. *Chemical Physics Letters*, 404(1-3):13–18, March 2005.

[54] Saumil J Gandhi, Daniel Zenklusen, Timothée Lionnet, and Robert H Singer. Transcription of functionally related constitutive genes is not coordinated. *Nature structural & molecular biology*, 18(1):27–34, January 2011.

[55] Nathalie (Leica Microsystems) Garin. Comparison of Deconvolution Software - Part 2. *G.I.T Imaging & Microscopy*, 3:2–4, 2010.

[56] Ethan C Garner. MicrobeTracker: quantitative image analysis designed for the smallest organisms. *Molecular microbiology*, 80(3):577–9, May 2011.

[57] Sarah Geisler, Lisa Lojek, Ahmad M Khalil, Kristian E Baker, and Jeff Coller. Decapping of Long Noncoding RNAs Regulates Inducible Genes. *Molecular cell*, 45(3):279–291, January 2012.

[58] Sina Ghaemmaghami, Won-Ki Huh, Kiowa Bower, Russell W Howson, Archana Belle, Noah Dephoure, Erin K O'Shea, and Jonathan S Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959): 737–41, October 2003.

[59] Sina Ghaemmaghami, Won-Ki Huh, Kiowa Bower, Russell W Howson, Archana Belle, Noah Dephoure, Erin K O'Shea, and Jonathan S Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959): 737–41, October 2003.

[60] Sina Ghaemmaghami, Won-Ki Huh, Kiowa Bower, Russell W Howson, Archana Belle, Noah Dephoure, Erin K O'Shea, and Jonathan S Weiss-

man. Global analysis of protein expression in yeast. *Nature*, 425(6959): 737–41, October 2003.

[61] RI Ghauharali and GJ Brakenhoff. Fluorescence photobleaching-based image standardization for fluorescence microscopy. *Journal of Microscopy*, (April 1999):88–100, 2000.

[62] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, December 2005.

[63] a. Golightly and D. J. Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820, September 2011.

[64] Paul C Goodwin. Evaluating optical aberration using fluorescent microspheres: methods, analysis, and corrective actions. *Methods in cell biology*, 81(06):397–413, January 2007.

[65] Andrew Gordon, Alejandro Colman-Lerner, TE Tina E Chin, Kirsten R Benjamin, Richard C Yu, and Roger Brent. Single-cell quantification of molecules and rates using open-source microscope-based cytometry. *Nature methods*, 4(2):175–181, 2007.

[66] Danna Gurari, Diane Theriault, Mehrnoosh Sameki, Brett Isenberg, Tuan a. Pham, Alberto Purwada, Patricia Solski, Matthew Walker, Chentian Zhang, Joyce Y. Wong, and Margrit Betke. How to Collect Segmentations for Biomedical Images? A Benchmark Evaluating the Performance of Experts, Crowdsourced Non-experts, and Algorithms. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 1169–1176. IEEE, January 2015.

[67] Cheol Woong Ha and Won-Ki Huh. The implication of Sir2 in replicative aging and senescence in Saccharomyces cerevisiae. *Aging*, 3(3):319–24, March 2011.

[68] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, December 2006.

[69] Elizabeth a Hackett, R Keith Esch, Seth Maleri, and Beverly Errede. A family of destabilized cyan fluorescent proteins as transcriptional reporters in S. cerevisiae. *Yeast (Chichester, England)*, 23(5):333–49, April 2006.

[70] Steven Hahn and Elton T Young. Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*, 189 (3):705–36, November 2011.

[71] Jeffrey E. Halley, Tommy Kaplan, Alice Y. Wang, Michael S. Kobor, and Jasper Rine. Roles for H2A.Z and its acetylation in GAL1 transcription and gene induction, but not GAL1-transcriptional memory. *PLoS Biology*, 8(6), 2010.

[72] Anders S Hansen and Erin K O'Shea. Promoter decoding of transcription factor dynamics involves a trade-off between noise and control of gene expression. *Molecular systems biology*, 9(704):704, 2013.

[73] Nan Hao and Erin K O'Shea. Signal-dependent dynamics of transcription factor translocation controls gene expression. *Nature structural & molecular biology*, 19(1):31–9, January 2012.

[74] Claire V Harper, Bärbel Finkenstädt, Dan J Woodcock, Sönke Friedrichsen, Sabrina Semprini, Louise Ashall, David G Spiller, John J Mullins, David a

Rand, Julian R E Davis, and Michael R H White. Dynamic analysis of stochastic transcription cycles. *PLoS biology*, 9(4):e1000607, April 2011.

[75] Marti a. Hearst, Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Schölkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

[76] Daniel Hebenstreit. Are gene loops the cause of transcriptional noise? *Trends in genetics : TIG*, 29(6):333–338, May 2013.

[77] R Heim, AB Cubitt, and RY Tsien. Improved green fluorescence. *Nature*, 373:663–664, 1995.

[78] Melanie Herscovitch, Eric Perkins, Andy Baltus, and Melina Fan. Addgene provides an open forum for plasmid sharing. *Nature Biotechnology*, 30(4): 316–317, 2012.

[79] Pascal Hersen, Megan N McClean, L Mahadevan, and Sharad Ramanathan. Signal processing by the HOG MAP kinase pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 105(20): 7165–70, May 2008.

[80] Keng Imm Hng and Dirk Dormann. ConfocalCheck–a software tool for the automated monitoring of confocal microscope performance. *PloS one*, 8 (11):e79879, 2013.

[81] Sami Hocine, Pascal Raymond, Daniel Zenklusen, Jeffrey A Chao, and Robert H Singer. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nature methods*, 10(2):119–21, 2013.

[82] J Horak and D H Wolf. Catabolite inactivation of the galactose transporter in the yeast Saccharomyces cerevisiae : ubiquitination , endocytosis , and

Catabolite Inactivation of the Galactose Transporter in the Yeast Saccharomyces cerevisiae : Ubiquitination , Endocytosis , and. 762(41), 1997.

[83] Gil Hornung, Raz Bar-Ziv, Dalia Rosin, Nobuhiko Tokuriki, Dan S. Tawfik, Moshe Oren, and Naama Barkai. Noise-mean relationship in mutated promoters. *Genome Research*, July 2012.

[84] Jonathan Houseley, Liudmilla Rubbi, Michael Grunstein, David Tollervey, and Maria Vogelauer. A ncRNA Modulates Histone Modification and mRNA Induction in the Yeast GAL Gene Cluster. *Molecular Cell*, 32(5): 685–695, December 2008.

[85] John R. Houser, Eintou Ford, Sudeshna M. Chatterjea, Seth Maleri, Timothy C. Elston, and Beverly Errede. An improved short-lived fluorescent protein transcriptional reporter for Saccharomyces cerevisiae. *Yeast*, 29(12): 519–530, December 2012.

[86] Chieh Hsu, Simone Scherrer, Antoine Buetti-Dinh, Prasuna Ratna, Julia Pizzolato, Vincent Jaquet, and Attila Becskei. Stochastic signalling rewires the interaction map of a multiple feedback network during yeast evolution. *Nature communications*, 3:682, January 2012.

[87] Won-Ki Huh, James V Falvo, Luke C Gerke, Adam S Carroll, Russell W Howson, Jonathan S Weissman, and Erin K O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–91, October 2003.

[88] T Ideker, V Thorsson, J a Ranish, R Christmas, J Buhler, J K Eng, R Bumgarner, D R Goodlett, R Aebersold, and L Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (New York, N.Y.)*, 292(5518):929–934, 2001.

[89] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, 44(1):87–116, 1988.

[90] P. Jayadeva Bhat, D. Oh, and J. E. Hopper. Analysis of the GAL3 signal transduction pathway activating GAL4 protein-dependent transcription in Saccharomyces cerevisiae. *Genetics*, 125(2):281–291, 1990.

[91] M Johnston, J S Flick, and T Pexton. Multiple mechanisms provide rapid and stringent glucose repression of GAL gene expression in Saccharomyces cerevisiae. *Molecular and cellular biology*, 14(6):3834–41, June 1994.

[92] S Kalies, K Kuetemeyer, and a Heisterkamp. Mechanisms of high-order photobleaching and its relationship to intracellular ablation. *Biomedical optics express*, 2(4):805–16, January 2011.

[93] Rajesh Kumar Kar, Mohd Tanvir Qureshi, Akshay Kumar Dasadhikari, Taiyeb Zahir, Kareenhalli V. Venkatesh, and Paike Jayadeva Bhat. Stochastic galactokinase expression underlies GAL gene induction in a GAL3 mutant of Saccharomyces cerevisiae. *FEBS Journal*, 281(7):1798–1817, 2014.

[94] Jonathan R. Karr, Jayodita C. Sanghvi, Derek N. Macklin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I. Glass, and Markus W. Covert. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*, 150(2):389–401, July 2012.

[95] Benjamin B Kaufmann and Alexander van Oudenaarden. Stochastic gene expression: from single molecules to the proteome. *Current opinion in genetics & development*, 17(2):107–12, April 2007.

[96] Benjamin B. Kaufmann, Qiong Yang, Jerome T. Mettetal, and Alexander Van Oudenaarden. Heritable stochastic switching revealed by single-cell genealogy. *PLoS Biology*, 5(9):1973–1980, 2007.

[97] H. Kempe, a. Schwabe, F. Cremazy, P. J. Verschure, and F. J. Bruggeman. The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. *Molecular Biology of the Cell*, pages 797–804, 2014.

[98] Birgit Kraus, Manja Ziegler, and Horst Wolff. Linear fluorescence unmixing in cell biological research. *Modern research and educational topics in microscopy*, pages 863–872, 2007.

[99] Sharmistha Kundu and Craig L. Peterson. Role of chromatin states in transcriptional memory. *Biochimica et Biophysica Acta - General Subjects*, 1790(6):445–455, 2009.

[100] Sharmistha Kundu and Craig L Peterson. Dominant role for signal transduction in the transcriptional memory of yeast GAL genes. *Molecular and cellular biology*, 30(10):2330–2340, 2010.

[101] Mats Kvarnström, Katarina Logg, Alfredo Diez, Kristofer Bodvard, and Mikael Käll. Image analysis algorithms for cell contour recognition in budding yeast. *Optics express*, 16(17):12943–57, August 2008.

[102] Amy J. Lam, François St-Pierre, Yiyang Gong, Jesse D. Marshall, Paula J. Cranfill, Michelle a. Baird, Michael R. McKeown, Jörg Wiedenmann, Michael W. Davidson, Mark J. Schnitzer, Roger Y. Tsien, and Michael Z. Lin. Improving FRET dynamic range with bright green and red fluorescent proteins. (September), 2012.

[103] Philip J. Lee, Noah C. Helman, Wendell a. Lim, and Paul J. Hung. A microfluidic system for dynamic yeast cell imaging. *BioTechniques*, 44(1): 91–95, 2008.

[104] Sidae Lee, Wendell a Lim, and Kurt S Thorn. Improved Blue, Green, and

Red Fluorescent Protein Tagging Vectors for S. cerevisiae. *PloS one*, 8(7): e67902, January 2013.

[105] Bing Li, Samantha G Pattenden, Daeyoup Lee, José Gutiérrez, Jie Chen, Chris Seidel, Jennifer Gerton, and Jerry L Workman. Preferential occupancy of histone variant H2AZ at inactive promoters influences local histone modifications and chromatin remodeling. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18385–18390, 2005.

[106] Catherine a Lichten, Rachel White, Ivan Bn Clark, and Peter S Swain. Unmixing of fluorescence spectra to resolve quantitative time-series measurements of gene expression in plate readers. *BMC biotechnology*, 14(1): 11, 2014.

[107] Jennifer Lippincott-Schwartz and George H Patterson. Development and use of fluorescent protein markers in living cells. *Science (New York, N.Y.)*, 300(5616):87–91, 2003.

[108] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9 (7):637–637, June 2012.

[109] J.C.W. Locke and M.B. Elowitz. Using Movies to Analyse Gene Circuit Dynamics in Single Cells. *Nature Reviews Microbiology*, 7(5):383–392, 2009.

[110] D Lohr, P Venkov, and J Zlatanova. Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 9(9):777–787, 1995.

[111] Peng Lu, Christine Vogel, Rong Wang, Xin Yao, and Edward M Marcotte. Absolute protein expression profiling estimates the relative contributions of

transcriptional and translational regulation. *Nature biotechnology*, 25(1): 117–24, January 2007.

[112] David J C Mackay. *Information Theory , Inference , and Learning Algorithms*. Cambridge University Press, 2003.

[113] Celine I Maeder, Mark a Hink, Ali Kinkhabwala, Reinhard Mayr, Philippe I H Bastiaens, and Michael Knop. Spatial regulation of Fus3 MAP kinase activity through a reaction-diffusion mechanism in yeast pheromone signalling. *Nature cell biology*, 9(11):1319–26, November 2007.

[114] Amir Massoudi, Dimitri Semenovich, and Arcot Sowmya. Cell tracking and mitosis detection using splitting flow networks in phase-contrast imaging. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2012:5310–3, August 2012.

[115] J Cooper Mcdonald, David C Duffy, Janelle R Anderson, and Daniel T Chiu. Review - Fabrication of microfluidic systems in poly (dimethylsiloxane). *Electrophoresis*, 21:27–40, 2000.

[116] T McInerney, T McInerney, D Terzopoulos, and D Terzopoulos. Deformable models in medical image analysis: a survey. *Medical image analysis*, 1(2): 91–108, 1996.

[117] Erik Meijering. Cell Segmentation: 50 Years Down the Road [Life Sciences]. *IEEE Signal Processing Magazine*, 29(5):140–145, September 2012.

[118] Jerome T Mettetal, Dale Muzzey, Carlos Gómez-Uribe, and Alexander van Oudenaarden. The frequency dependence of osmo-adaptation in Saccharomyces cerevisiae. *Science (New York, N.Y.)*, 319(5862):482–484, 2008.

[119] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.

[120] Nacho Molina, David M Suter, Rosamaria Cannavo, Benjamin Zoller, Ivana Gotic, and Félix Naef. Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proceedings of the National Academy of Sciences of the United States of America*, 110(51):20563–8, 2013.

[121] Butch Moomaw. Camera technologies for low light imaging: overview and relative advantages. *Methods in cell biology*, 81(06):251–83, January 2007.

[122] B. Munsky, G. Neuert, and A. van Oudenaarden. Using Gene Expression Noise to Understand Gene Regulation. *Science*, 336(6078):183–187, April 2012.

[123] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular systems biology*, 5(2001):318, 2009.

[124] Chitra R Nayak and Andrew D Rutenberg. Quantification of fluorophore copy number from intrinsic fluctuations during fluorescence photobleaching. *Biophysical journal*, 101(9):2284–93, November 2011.

[125] G. Neuert, B. Munsky, R. Z. Tan, L. Teytelman, M. Khammash, and a. van Oudenaarden. Systematic Identification of Signal-Activated Stochastic Gene Regulation. *Science*, 339(6119):584–587, January 2013.

[126] John R S Newman, Sina Ghaemmaghami, Jan Ihmels, David K Breslow, Matthew Noble, Joseph L DeRisi, and Jonathan S Weissman. Single-cell

proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–6, June 2006.

[127] Feng Ning, Damien Delhomme, Yann LeCun, Fabio Piano, Léon Bottou, and Paolo Emilio Barbano. Toward automatic phenotyping of developing embryos from videos. *IEEE Transactions on Image Processing*, 14(9):1360–1371, 2005.

[128] Leah M Octavio, Kamil Gedeon, and Narendra Maheshri. Epigenetic and conventional regulation is distributed among activators of FLO11 allowing tuning of population-level heterogeneity in its expression. *PLoS genetics*, 5 (10):e1000673, October 2009.

[129] Burak Okumus, Sadik Yildiz, and Erdal Toprak. Fluidic and microfluidic tools for quantitative systems biology. *Current Opinion in Biotechnology*, 25:30–38, 2014.

[130] Jamie Owen, Darren J. Wilkinson, and Colin S. Gillespie. Likelihood free inference for Markov processes: a comparison. *ArXiv e-prints*, pages 1–25, 2014.

[131] Ertugrul M Ozbudak, Mukund Thattai, Iren Kurtser, Alan D Grossman, and Alexander van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature genetics*, 31(1):69–73, May 2002.

[132] Olivia Padovan-Merhar, Gautham P. Nair, Andrew G. Biaesch, Andreas Mayer, Steven Scarfone, Shawn W. Foley, Angela R. Wu, L. Stirling Churchman, Abhyudai Singh, and Arjun Raj. Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. *Molecular Cell*, pages 1–14, 2015.

[133] Saurabh Paliwal, Pablo a Iglesias, Kyle Campbell, Zoe Hilioti, Alex Groisman, and Andre Levchenko. MAPK-mediated bimodal gene expression and adaptive gradient sensing in yeast. *Nature*, 446(7131):46–51, 2007.

[134] G H Patterson, S M Knobel, W D Sharif, S R Kain, and D W Piston. Use of the green fluorescent protein and its mutants in quantitative fluorescence microscopy. *Biophysical journal*, 73(November):2782–2790, 1997.

[135] Johan Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175, 2005.

[136] Jean-Denis Pédelacq, Stéphanie Cabantous, Timothy Tran, Thomas C Terwilliger, and Geoffrey S Waldo. Engineering and characterization of a superfolder green fluorescent protein. *Nature biotechnology*, 24(1):79–88, January 2006.

[137] Vicent Pelechano and Lars M Steinmetz. Gene regulation by antisense transcription. *Nature reviews. Genetics*, 14(12):880–93, December 2013.

[138] Serge Pelet, Reinhard Dechant, Sung Sik Lee, Frank van Drogen, and Matthias Peter. An integrated image analysis platform to quantify signal transduction in single cells. *Integrative biology : quantitative biosciences from nano to macro*, (207890), September 2012.

[139] Theodore J Perkins, Andrea Y Weisse, and Peter S Swain. Chance and Memory. *Quantitative Biology*, pages 51–72, 2013.

[140] Marina Pinskaya, Stéphanie Gourvennec, and Antonin Morillon. H3 lysine 4 di- and tri-methylation deposited by cryptic transcription attenuates promoter activation. *The EMBO Journal*, 28(12):1697–1707, June 2009.

[141] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance:

stochastic gene expression and its consequences. *Cell*, 135(2):216–26, October 2008.

[142] Arjun Raj, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):1707–1719, 2006.

[143] Oliver J Rando and Fred Winston. Chromatin and transcription in yeast. *Genetics*, 190(2):351–87, February 2012.

[144] Jonathan M Raser and Erin K O'Shea. Control of stochasticity in eukaryotic gene expression. *Science (New York, N.Y.)*, 304(5678):1811–4, June 2004.

[145] Ivan Rasnik, Todd French, Ken Jacobson, and Keith Berland. Electronic cameras for low-light microscopy. *Methods in cell biology*, 81(06):219–49, January 2007.

[146] B G Reid and G C Flynn. Chromophore formation in green fluorescent protein. *Biochemistry*, 36(22):6786–91, June 1997.

[147] Wolfgang Rettig, Bernd Strehmel, Sigurd Schrader, and Holger Seifert. *Applied Fluorescence in Chemistry, Biology and Medicine*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[148] Nitzan Rosenfeld, Jonathan W Young, Uri Alon, Peter S Swain, and Michael B Elowitz. Gene regulation at the single-cell level. *Science (New York, N.Y.)*, 307(5717):1962–5, March 2005.

[149] Nitzan Rosenfeld, Theodore J Perkins, Uri Alon, Michael B Elowitz, and Peter S Swain. A fluctuation method to quantify in vivo fluorescence data. *Biophysical journal*, 91(2):759–66, July 2006.

[150] James Ryley and Olivia M Pereira-Smith. Microfluidics device for single

cell gene expression analysis in Saccharomyces cerevisiae. *Yeast (Chichester, England)*, 23(14-15):1065–73, January 2006.

[151] Kevin R. Sanft, Sheng Wu, Min Roh, Jin Fu, Rone Kwei Lim, and Linda R. Petzold. StochKit2: Software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17):2457–2458, 2011.

[152] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Validity conditions for stochastic chemical kinetics in diffusion-limited systems. *Journal of Chemical Physics*, 140(5):054111, February 2014.

[153] Nathan C Shaner, Gerard G Lambert, Andrew Chammas, Yuhui Ni, Paula J Cranfill, Michelle a Baird, Brittney R Sell, John R Allen, Richard N Day, Maria Israelsson, Michael W Davidson, and Jiwu Wang. A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum. *Nature methods*, 10(5):407–9, May 2013.

[154] Dmitry Shcherbo, Christopher S Murphy, Galina V Ermakova, Elena a Solovieva, Tatiana V Chepurnykh, Aleksandr S Shcheglov, Vladislav V Verkhusha, Vladimir Z Pletnev, Kristin L Hazelwood, Patrick M Roche, Sergey Lukyanov, Andrey G Zaraisky, Michael W Davidson, and Dmitriy M Chudakov. Far-red fluorescent tags for protein imaging in living tissues. *The Biochemical journal*, 418(3):567–574, 2009.

[155] Sung Sik, Ima Avalos, Daphne H E W Huberts, Luke P Lee, and Matthias Heinemann. Whole lifespan microscopic observation of budding yeast aging through a micro fl uidic dissection platform. *PNAS*, 109(13):4916 – 4920, 2012.

[156] Robert S Sikorski and Philip Hieter. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics*, 122(1):19–27, 1989.

[157] Daniel Sinnecker, Philipp Voigt, Nicole Hellwig, and Michael Schaefer. Reversible photobleaching of enhanced green fluorescent proteins. *Biochemistry*, 44(18):7085–94, May 2005.

[158] Ron Skupsky, John C. Burnett, Jonathan E. Foley, David V. Schaffer, and Adam P. Arkin. HIV promoter integration site primarily modulates transcriptional burst size rather than frequency. *PLoS Computational Biology*, 6(9), 2010.

[159] Christian J. Slubowski, Alyssa D. Funk, Joseph M. Roesner, Scott M. Paulissen, and Linda S. Huang. Plasmids for C-terminal tagging in Saccharomyces cerevisiae that contain improved GFP proteins, Envy and Ivy. *Yeast*, 32(4):379–387, April 2015.

[160] Greenfield Sluder and Joshua J Nordberg. Microscope basics. *Methods in cell biology*, 81(06):1–10, January 2007.

[161] Ec Small, L Xi, and Jp Wang. Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Pnas*, 111(24):E2462–71, 2014.

[162] Lok-Hang So, Anandamohan Ghosh, Chenghang Zong, Leonardo a Sepúlveda, Ronen Segev, and Ido Golding. General properties of transcriptional time series in Escherichia coli. *Nature genetics*, 43(6):554–60, June 2011.

[163] Carl Song, Hilary Phenix, Vida Abedi, Matthew Scott, Brian P. Ingalls, Mads Kæ rn, and Theodore J. Perkins. Estimating the stochastic bifurcation structure of cellular networks. *PLoS Computational Biology*, 6(3), 2010.

[164] L Song, R P van Gijlswijk, I T Young, and H J Tanke. Influence of fluo-

rochrome labeling density on the photobleaching kinetics of fluorescein in microscopy. *Cytometry*, 27(3):213–23, March 1997.

[165] Loling Song, E J Hennink, Ted Young, and Hans J Tanke. Photobleaching kinetics of fluorescein in quantitative fluorescence microscopy. *Biophysical Journal*, (June 1995):2588–2600, 1995.

[166] Kenneth R Spring. Cameras for digital microscopy. *Methods in cell biology*, 81(06):171–86, January 2007.

[167] Jacob Stewart-Ornstein, Jonathan S. Weissman, and Hana El-Samad. Cellular Noise Regulons Underlie Fluctuations in Saccharomyces cerevisiae. *Molecular Cell*, 45(4):483–493, February 2012.

[168] Sarah R. Stockwell, Christian R. Landry, and Scott a. Rifkin. The yeast galactose network as a quantitative model for cellular memory. *Mol. BioSyst.*, 11(1):28–37, 2015.

[169] Aaron F Straight. Fluorescent protein applications in microscopy. *Methods in cell biology*, 81(06):93–113, January 2007.

[170] Atsushi Suenaga, Noriaki Okimoto, Noriyuki Futatsugi, Yoshinori Hirano, Tetsu Narumi, Yousuke Ohno, Ryoko Yanai, Takatsugu Hirokawa, Toshikazu Ebisuzaki, Akihiko Konagaya, and Makoto Taiji. Structure and dynamics of RNA polymerase II elongation complex. *Biochemical and Biophysical Research Communications*, 343(1):90–98, 2006.

[171] Kevin Francis Sullivan and Steve A Kay. *Green Fluorescent Proteins*. Gulf Professional Publishing, 1999.

[172] D.M. David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *science*, 332(6028):472–4, April 2011.

[173] Peter S Swain, Michael B Elowitz, and Eric D Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20): 12795–800, October 2002.

[174] R J Taylor, D Falconnet, a Niemistö, S a Ramsey, S Prinz, I Shmulevich, T Galitski, and C L Hansen. Dynamic analysis of MAPK signaling using a high-throughput microfluidic single-cell imaging platform. *Proceedings of the National Academy of Sciences of the United States of America*, 106(10): 3758–3763, 2009.

[175] Shu-Wen Teng, Yufang Wang, Kimberly C Tu, Tao Long, Pankaj Mehta, Ned S Wingreen, Bonnie L Bassler, and N P Ong. Measurement of the copy number of the master quorum-sensing regulator of a bacterial cell. *Biophysical journal*, 98(9):2024–31, May 2010.

[176] Marc Tramier, Morad Zahid, Jean-claude Mevel, and Marie-jo Masse. Sensitivity of CFP / YFP and GFP / mCherry Pairs to Donor Photobleaching on FRET Determination by Fluorescence Lifetime Imaging Microscopy in Living Cells. 939(August):933–939, 2006.

[177] Ana Traven, Branka Jelicic, and Mary Sopta. Yeast Gal4: a transcriptional paradigm revisited. *EMBO reports*, 7(5):496–499, 2006.

[178] Roger Y Tsien. the Green Fluorescent. *Annu. Rev. Biochem.*, 67:509 – 544, 1998.

[179] G Ullman, M Wallden, E G Marklund, A Mahmutovic, Ivan Razinkov, and J Elf. High-throughput gene expression analysis at the level of single proteins using a microfluidic turbidostat and automated cell tracking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1611):20120025, February 2013.

[180] Dmitri Volfson, Jennifer Marciniak, William J Blake, Natalie Ostroff, Lev S Tsimring, and Jeff Hasty. Origins of extrinsic variability in eukaryotic gene expression. *Nature*, 439(7078):861–4, February 2006.

[181] C Vonesch and M Unser. A fast thresholded landweber algorithm for wavelet-regularized multidimensional deconvolution. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 17 (4):539–49, April 2008.

[182] Yuichi Wakamoto, Neeraj Dhar, Remy Chait, Katrin Schneider, François Signorino-Gelo, Stanislas Leibler, and John D McKinney. Dynamic persistence of antibiotic-stressed mycobacteria. *Science (New York, N.Y.)*, 339 (6115):91–5, January 2013.

[183] Quanli Wang, Jarad Niemi, Chee-Meng Meng Tan, Lingchong You, and Mike West. Image segmentation and dynamic lineage analysis in single-cell fluorescence microscopy. *Cytometry Part A*, 77(1):101–110, January 2010.

[184] Xiao Wang, Beverly Errede, and Timothy C Elston. Mathematical analysis and quantification of fluorescent proteins as transcriptional reporters. *Biophysical journal*, 94(6):2017–26, March 2008.

[185] Yu-Li Wang. Computational restoration of fluorescence images: noise reduction, deconvolution, and pattern recognition. *Methods in cell biology*, 81(06):435–45, January 2007.

[186] Jennifer C Waters. Live-cell fluorescence imaging. *Methods in cell biology*, 81(06):115–40, January 2007.

[187] Jennifer C Waters. Accuracy and precision in quantitative fluorescence microscopy. *The Journal of cell biology*, 185(7):1135–48, June 2009.

[188] Jennifer C Waters and Jason R Swedlow. Techniques Interpreting Fluorescence Microscopy Images and Measurements. In *Evaluating Techniques in Biochemical Research*, pages 36–42. 2008.

[189] Leehee Weinberger, Yoav Voichek, Itay Tirosh, Gil Hornung, Ido Amit, and Naama Barkai. Expression Noise and Acetylation Profiles Distinguish HDAC Functions. *Molecular cell*, 47(2):193–202, June 2012.

[190] Iestyn Whitehouse, Oliver J Rando, Jeff Delrow, and Toshio Tsukiyama. Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, 450(7172):1031–5, December 2007.

[191] Darren J Wilkinson. Parameter inference for stochastic kinetic models of bacterial gene regulation : a Bayesian approach to systems biology. In *Bayesian Statistics 9*, number Wilkinson 2009, pages 679–706. 2010.

[192] Jian-Qiu Wu, Chad D McCormick, and Thomas D Pollard. *Chapter 9: Counting proteins in living cells by quantitative fluorescence microscopy with internal standards.*, volume 89. Elsevier Inc., 1 edition, January 2008.

[193] Samuel Zambrano, Marco E Bianchi, Alessandra Agresti, and Nacho Molina. Stochasticity and negative feedback lead to pulsed dynamics and distinct gene activity patterns. Technical report, December 2014.

[194] Christoph Zechner and Heinz Koeppl. Uncoupled Analysis of Stochastic Reaction Networks in Fluctuating Environments. (December):7, 2014.

[195] Christoph Zechner, Michael Unger, Serge Pelet, Matthias Peter, and Heinz Koeppl. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature Methods*, 11(2):197–202, January 2014.

[196] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-RNA

counting reveals alternative modes of gene expression in yeast. *Nature structural & molecular biology*, 15(12):1263–71, December 2008.

[197] Chengda Zhang and James B. Konopka. A photostable green fluorescent protein variant for analysis of protein localization in Candida albicans. *Eukaryotic Cell*, 9(1):224–226, 2010.

[198] NL Zhang and David Poole. Exploiting causal independence in Bayesian network inference. *Journal of Artificial Intelligence Research*, 5:301–328, 1996.

[199] C. J. Zopf, Katie Quinn, Joshua Zeidman, and Narendra Maheshri. Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression. *PLoS Computational Biology*, 9(7):e1003161, July 2013.

[200] Christopher J. Zopf and Narendra Maheshri. Acquiring Fluorescence Time-lapse Movies of Budding Yeast and Analyzing Single-cell Dynamics using GRAFTS. *Journal of Visualized Experiments*, (77), July 2013.

Appendix A

Appendix to Microscope Characterisation

A.1 Protocol for the Calculation of Flat Field Correction Using the Microfluidic device

The following is a detailed protocol for the calculation of Flat field correction.

1. Prepare a single chamber ALCATRAS device according to the standard lab protocol. Note that only one inlet hole is required.
2. Prepare a 5ml syringe with filter sterilised (0.22 μm) fluorescent dye. The dye and concentration to be used is channel dependent, and for GFP or GFPAutoFL 0.001% (mass by volume) fluorescein was used.
3. mount the device on the slide holder following the standard lab protocol.
4. Connect PTFE tubing to the fluorescein syringe and mount in the syringe pump in the usual way.

5. Squeezing by hand, push fluorescein through the device.

6. Set the pump at μl/minute, secure the slide holder and tubes and leave for 2 hours to equilibriate.

7. Using the automated microscopy control software, set an acquisition of at least 50 positions for 50 time points, keeping the perfect focus system (PFS) on. The positions need not be independent, and in our case we would set three sets of tiled positions ofset by approximately 5μm (half a trap width).

8. Check images for reasonable temporal consistency using imageJ. If there is a significant drift in the mean over time then it may be necessary to leave the device to equilibriate for longer.

9. I images are satisfactory run the `BackgroundCorrectionRun.m` script in the `BackgroundCorrection` package. The image generated is a normalised mean intensity, and the inverse is used for flat field correction.

The same protocol was used for acquiring noise images, but the ALCATRAS device was replaced with a y-channel device and the flow set to 20 μl/minute. To acquire wide ranges of pixel values, a field of view was selected outside the channel but close to the edge, and defocused.

A.2 Measuring Camera Noise

A.2.1 Proof of sample-measurement linearity for CCD camera

We assume that exposure time is a proxy for sample brightness, so that if the same sample is exposed for $t_1$ms and t_2 ms the sample brightness in the second

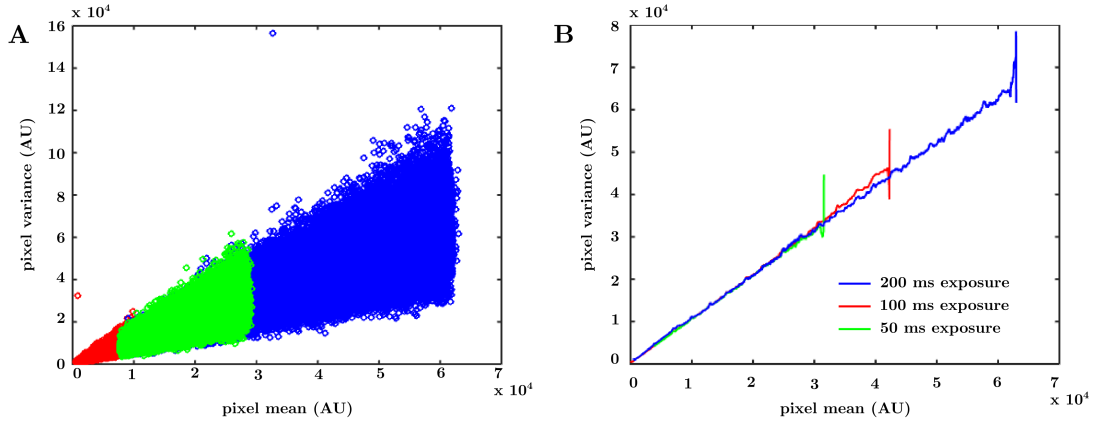

Figure A.1: Construction of mean-variance relationship for different camera settings and the independence of pixel variance from exposure time. Panel A shows the construction of the mean variance relationship. Each point is the mean and variance calculated for an individual pixel over fifty images of the same field of view, with the different colours indicating different fields of view used to cover the full range of pixel values for a given camera settings (in this case 200 ms exposure with CCD mode of the camera). Panel B shows smoothed mean and variance relationships for different exposure times. These mean-variance relationships are calculated by taking data like that shown in panel A and averaging over a 1000 value window to give a reasonably smooth plot. Mean-variance relationships are shown for 200 ms exposure (blue), 100 ms exposure (red) and 50 ms exposure (green) using the CCD mode of the camera. Clearly the exposure time has no significant affect on mean-variance relationship, and the linearity indicates that the noise is dominated by shot noise as one would expect for a modern, cooled CCD camera.

case, x_2, will obey the relationship $x_2 = x_1 \times (t_2/t_1)$.

We want to know the relationship between measurement y and sample brightness x, and particularly if this relationship is linear. Define this relationship as $y = f(x)$. If two measurements, y_1 and y_2 are made of two samples known to be related by the relationship $x_2 = nx_1$ then:

$$\frac{y_2}{y_1} = \frac{f(nx_1)}{f(x_1)}$$

if $y_2/y_1 = n$ for all n, then $f(nx_1)/f(x_1) = n$ for all n and f is a linear function. In figure 2.4 we have shown this for a large range of x values and n's, and so I would say we have shown satisfactorily that the function f relating sample brightness and measured intensity is linear for the CCD mode of the camera.

Appendix B

Appendix to Fluorophore Selection and Characterisation

B.1 Production of UBI-MΔk-FLUOR plasmids

Plasmids were produced using the clonetec In-Fusion®system with the Thorn lab [104] plasmids pFA6-GFPgamma-SpHis5 and pFA6-mCherry-CaURA3 plasmids as a backbone. The In-Fusion® can simultaneously ligate up to five DNA fragments. Fragments can be PCR amplified with primers selected for homology between fragments to be joined and homology with the plasmid backbone for the first and last fragment; ligation is then a simple single step reaction with subsequent transformation and purification in ecoli. To produce the plasmids described pFA6-GFPgamma-SpHis5 and pFA6-mCherry-CaURA3 were digested with Pac1 and Mlu1 to excise the fluorophore while leaving the selection marker and plasmid backbone; this larger section was isolated by size on an agar gel and purified using a standard gel purification kit. mKate2 and GFPγ sequences were amplified from pFA6-GFPgamma-SpHis5 and pFA6-mKate-SpHis5 plasmids re-

spectively and the UBI-MΔk sequence amplified from the pnc1124 deposited with Addgene by Houser et al.. The In-Fusion system was used to ligate these three fragments (UBI-MΔk, fluorophore sequence and plasmid backbone) and the resulting plasmid mix was transformed into *E. coli* . Individual ampicilin resistant colonies were selected and their plasmid content purified by standard mini-prep kit before being tested by a range of digests to distinguish successful plasmids from parent plasmids. For each plasmid a colony deemed successful by the digest results was selected and its plasmid purified by midi-prep before being sequenced. Discrepancies between the sequenced plasmid and the published sequences were only found in the selection markers which subsequent transformations and selection proved to be inconsequential.

B.2 Details of the Measurement of Decay Rate by Fixation and Flow Cytometry

To measure the decay rate of the three fluorophores an experiment similar to that of Houser et al. was undertaken. The three strains were grown overnight in XY medium with raffinose (2%) and galactose (0.1%). In the morning cells were resuspended in fresh XY raffinose/galactose at an OD of 0.2. After 1 hour, the cells were spun down, washed with XY and resuspended to an OD of 0.2 in XY glucose (2%) media. From that time onwards, culture samples were collected at intervals of approximately 45 minutes and fixed using a standard paraformaldhyde fixing protocol detailed below. The time of each fixation was recorded as the time at which paraformaldhyde was applied. The fixed samples were analysed by flow cytometry, which allowed high throughput measurements of the large number of samples produced. Identical settings were used for each sample, with 100,000

events being obtained. Gating was applied after acquistion based on forward and side scatter.

B.3 Cell Fixation by Paraformaldhyde

Cells were fixed according to the following protocol provided by C. Josephides and adapted from Biggins protocol for yeast cell fixation (revision 1).

1. Spin cells (6,000 rpm for 1 minute) in eppendorfs and remove supernatant.
2. Add 100uL of 4% paraformaldehyde and vortex.
3. Incubate at RT for 15 minutes.
4. Spin cells (6,000 rpm for 1 minute) and remove supernatant.
5. Wash once in KPO4/sorbitol (0.75 mL).
6. Resuspend in 1ml of KPO4/sorbitol .
7. Store cells in refrigerator for up to one month.
8. Sonicate cells for 3 seconds (full power) before doing flow cytometry.

Appendix C

Appendix to Automated Segmentation

C.1 Algorithm 1: Matt's Algorithm

```
FIND TRAPS
DIC image 1 = get DIC image at timepoint 1
find trap locations at timepoint 1 by cross correlation of
 trap image with DIC image 1
[this defines a set of initial trap locations, one for
each trap in the image]
for t:={2 ... end}:
   DIC image t =  DIC image at timepoint t
   (x shift,y shift) = find image shift by cross
   correlation of DIC image 1 with DIC image t
   trap locations at timpoint t = traps locations at
   timepoint 1 + (x shift, y shift)
```

```
FIND CELLS
for t:={1 ... end}:
      DIC image t = DIC image at timepoint t
   for trap:=set of traps:
      trap decision image = classify DIC image t at
      location of trap
      [this produces a decision image: an image which has
      low values for pixels likely to be cell centres and
      high values elsewhere. This is acheived by applying
      a support vector machine to each pixel in a
      collection of transformed images]
      cell centre regions = divide into connected regions(
      trap decision image < decision image threshold)
      for cell := cell centre regions
         assign centre and radius to cell based on
         weighted average of hough transformed image over
         the cell region.

ASSIGN CELL LABELS
for trap:=set of traps:
   for cell:=cells in trap at timepoint 1
      assign new cell label to cell
for t:={2 ... end}:
   for trap:=set of traps:
      calculate modified euclidean distance between cells
      found at timepoint t and those found at timepoint (t
      -1)
```

```
[this modified distance takes includes change in
cell radius as a dimension]
d = minimum(euclidean distance between all pairs of
cells)
(cell A , cell B) = cells at timepoint t and (t-1)
respectively that are distance d apart
while d<distance threshold:
    set cell label of cell B to be cell label of cell
     A
    remove cell A and cell B from the set of cells to
     be assigned cell labels
    d = minimum(euclidean distance between all pairs
    of cells still awaiting cell labels)
assign new cell label to cells at timepoint t still
awaiting a cell label
```

C.2 Radial Gradient Transformation

The base image for transformation is generated by subtracting the one out of focus bright field image from the other. This is normalised by subtracting the median of the image and dividing it by the interquartile range to try and compensate illumination and focus differences between experiments. From this is calculated two gradient images:

$$\begin{aligned}\text{grad x}(x,y) &= \text{image}(x,y) - \text{image}(x-1,y)\\ \text{grad y}(x,y) &= \text{image}(x,y) - \text{image}(x,y-1)\end{aligned}$$

From these a final transformed image is calculated:

$$\begin{aligned}\theta(x,y) &= \arctan(\frac{y - \text{y cell centre}}{x - \text{x cell centre}})\\ \text{transformed image}(x,y) &= \text{grad x}(x,y) \times \cos(\theta(x,y)) + \text{grad y}(x,y) \times \sin(\theta(x,y))\end{aligned}$$

This produces an image that generally has high values at the cell edges, and is then inverted to provide low values.

C.3 Algorithm 3: Cross Correlation with Active Contour

As a result of casual inspection of the segmentation result on various data sets I felt that many of the errors could be rectified by making greater use of the information about the cell at timepoint t to identify the same cell (including successfully tracking it) at timepoint $t+1$. To try and effect this I implemented an algorithm inspired by the methods described in Blake and Isard [16] but without some of the efficiency increasing assumptions made therein. In outline, the algorithm attempts to take the outline of the cell at time point t and apply it to the image of the trap at time point $t+1$ to generate 'tracking image', which has high values for the centre of that cell. The construction of the tracking image has

been quite heuristic. I started using cross correlation, but found that this was dominated by rare bright spots and so was not very effective. I instead developed a method based on a modified circular hough transform which builds and accumulation array in the same way as the hough transform but only using edge pixels with the right direction of gradient. The result is an image that is very bright for the centres of white circles on a black background (of a specified radius), but dim for centres of a black circle on a white background. The directional hough transform of the time point t and image was calculated for all discrete radii, producing a vector of accumulation array values for each pixel. For cell c at time point t the raw tracking image for time point $t+1$ was calculated as the dot product of this accumulation array vector with the same accumulation array vector of the centre of cell c at time point t. The raw tracking image is then multiplied by a thresholded decision image to ensure only admissible cell centres are used, and then multiplied by a gaussian centred on the location of cell c at timepoint $t-1$ (to encode the fact that cells move very little from timepoint to timepoint) with a heavier weighting in the backwards direction to encode that cells tend to move further down the traps due to flow. The following is a description of the algorithm in pseudo code.

```
for timepoint=1:
        calculate decision image using cellVision model
        find new cells (decision image)

for timepoint={2 ... end}:
        calculate decision image using cellVision model
        for n = {1... number of cells identified at
           previous timepoint}
        calculate tracking image for cell n
        tracking image stack (:,:,n) = tracking image
```

```
        m = max(tracking image stack)
        while m > tracking image threshold
        (x,y,z) = location of m in tracking image stack
        add cell centre at x,y
        give the cell the tracking number of cell z at
        the previous timepoint
        find cell outline using active contour algorithm
        with outline from previous timepoint
        set cell area to (tracking image threshold)-1 in
        all slices of tracking image stack
        set cell area to (descision image threshold)+1
        in decision image
        m = max(modified tracking image stack)
        find new cells (modified decision image)

define: find new cells (decision image)
        m = max(decision image)
        while m < decision image threshold
                add cell centre at location of m in
                   decision image
                give the cell a new tracking number
                perform active contour algorithm to find
                   cell outline
                set new cell area to (decision image
                   threshold)+1
                in decision image
               m = max(altered decision image)
```

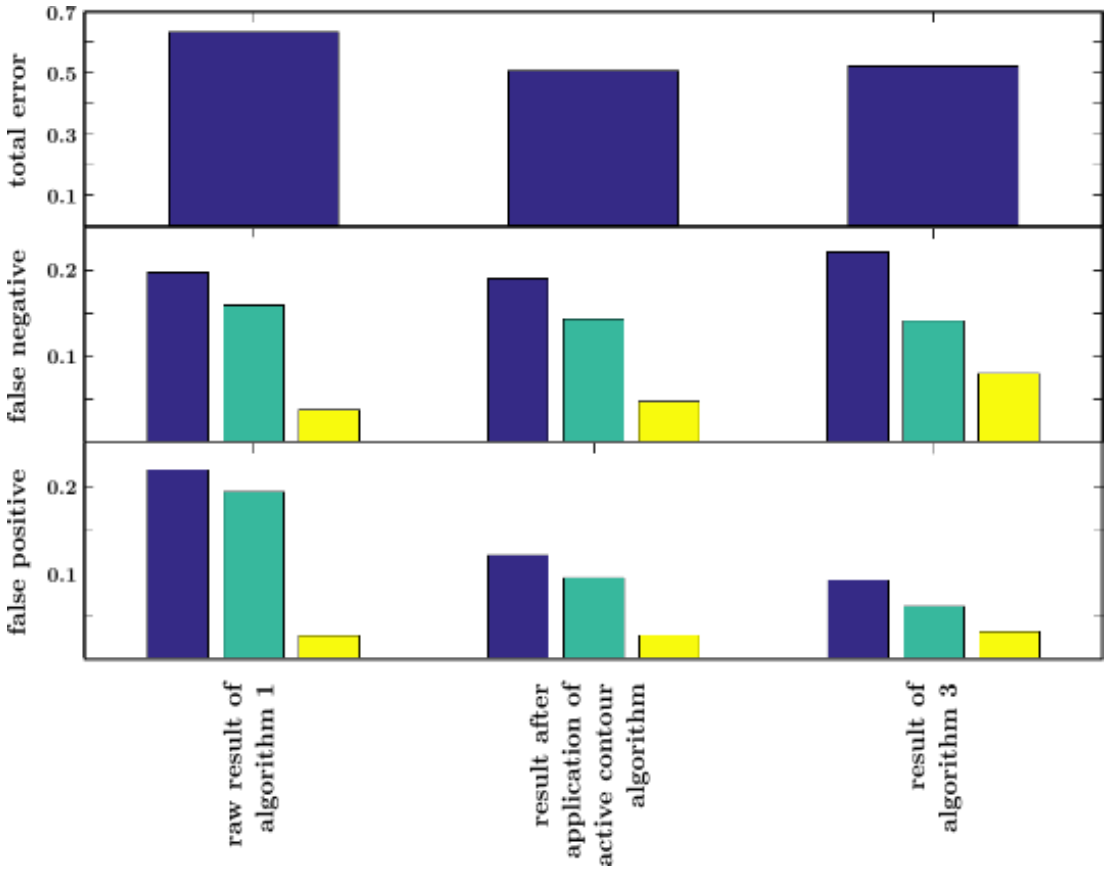

Figure C.1: Performance of algorithm 3(right) compared with the raw result of algorithm 1 (left) and the active contour the active contour method applied to algorithm 1 (centre). As can be seen, algorithm 3 is largely outperformed.

Results of Algorithm 3

A bar chart of the results for all three algorithms is plotted in figure C.1 (colours are the same as in figure 3.7 in chapter 3. It can be seen that algorithm 3 performs worse than it's predecessor. Manual inspection of the result seems to show that this is due to tracking errors in which a cell is mistakenly taken to be a new cell at the same location, which would explain the high rate of cell pixels assigned to the wrong cell (yellow bar in the second panel). This may mean that the algorithm will improve in performance if the active contour algorithm for edge detection is improved, or that a heuristic can correctly identify and merge these cases, but for now remains a work in progress.

Appendix D

Appendix to Estimation of Protein Concentration by Analysis of Stochastic Fluctuations in Photobleaching

D.1 Derivation of Modal Values for ν and p

As detailed in chapter 4, the posterior for ν and p without measurement error is given by:

$$p(\nu, p|\{I_i\}) = \left(\frac{1}{2\pi p(1-p)\nu}\right)^{\frac{M}{2}} \frac{1}{\prod_{i=0}^{M-1} I_i^{\frac{1}{2}}} \exp\left[-\frac{1}{2\nu p(1-p)} \sum_{i=0}^{M-1} \frac{(I_i p - I_{i+1})^2}{I_i}\right] \frac{p(\nu, p, I_0)}{p(\{I_i\})} \tag{D.1}$$

To calculate modal values for ν and p, we differentiatied the Likelihood $L =$

$p(\{I_i\}|\nu, p)$

$$\begin{aligned}
\ln L &= \frac{-M}{2}\ln(2\pi\nu pq) - \ln(\prod_{i=0}^{M-1} I_i^{\frac{1}{2}}) - \frac{1}{2\nu pq}\sum_{i=0}^{M-1}\frac{(I_i p - I_{i+1})^2}{I_i} \\
\frac{\partial(\ln L)}{\partial\nu} &= \frac{-M}{2}\frac{1}{\nu} + \frac{1}{2\nu^2 pq}\sum_{i=0}^{M-1}\frac{(I_i p - I_{i+1})^2}{I_i} \\
\frac{\partial(\ln L)}{\partial p} &= \frac{-M}{2pq}(1-2p) + \frac{1-2p}{2\nu p^2 q^2}\sum_{i=0}^{M-1}\frac{(I_i p - I_{i+1})^2}{I_i} + \frac{1}{\nu q}\sum_{i=0}^{M-1}(I_i p - I_{i+1})
\end{aligned} \tag{D.2}$$

where $q = p - 1$.

Setting the partial derivatives to zero, and using the notation

$$S = \sum_{i=0}^{M-1}\frac{(I_i p - I_{i+1})^2}{I_i}$$

we obtain:

$$\nu = \frac{S}{Mp(1-p)} \tag{D.3}$$

substituting this result into equation D.2 we obtain:

$$\begin{aligned}
0 &= \frac{-M}{2pq}(1-2p)\frac{S}{Mpq} + \frac{1-2p}{2p^2q^2}S + \frac{1}{q}\sum_{i=0}^{M-1}(I_i p - I_{i+1}) \\
0 &= -\frac{1-2p}{2p^2q^2}S + \frac{1-2p}{2p^2q^2}S + \frac{1}{q}\sum_{i=0}^{M-1}(I_i p - I_{i+1}) \\
0 &= \sum_{i=0}^{M-1}(I_i p - I_{i+1}) \\
p_{\text{modal}} &= \frac{\sum_{i=0}^{M-1} I_{i+1}}{\sum_{i=0}^{M-1} I_i}
\end{aligned}$$

Applying the proper notation:

$$p_{\text{modal}} = \frac{\sum_{i=0}^{M-1} I_{i+1}}{\sum_{i=0}^{M-1} I_i} \tag{D.4}$$

$$\nu_{\text{modal}} = \frac{S_{\text{modal}}}{Mp_{\text{modal}}(1-p_{\text{modal}})} \tag{D.5}$$

where:

$$S_{\text{modal}} = \sum_{i=0}^{M-1} \frac{(I_i p_{\text{modal}} - I_{i+1})^2}{I_i}$$

D.1.1 Derivation Expected Behaviour of Modal Value for ν

Due to the sum in the denominator, calculations of the statistical behaviour of our estimate of p is difficult, but if we assume that the data can determine p accurately, a claim that seems reasonable when we do not consider measurement error, then we can take p to be fixed and calculate the expected behaviour of ν_{modal}.

$$\begin{aligned} \langle \nu_{\text{modal}} \rangle &= \frac{\langle S \rangle}{Mp(1-p)} \\ \text{where } \langle S \rangle &= \langle \sum_{i=0}^{M-1} \frac{(I_i p - I_{i+1})^2}{I_i} \rangle \\ &= \nu_{\text{true}} \langle \sum_{i=0}^{M-1} \frac{(n_i p - n_{i+1})^2}{n_i} \rangle \end{aligned} \tag{D.6}$$

taking the first term of this series we have:

$$\begin{aligned}\langle\frac{(n_0p-n_1)^2}{n_0}\rangle &= \frac{\langle\langle n_1\rangle\rangle}{n_0}\\ &= \frac{n_0p(1-p)}{n_0}\\ &= p(1-p)\end{aligned}$$

looking at the second term we have:

$$\begin{aligned}\langle\frac{(n_1p-n_2)^2}{n_1}\rangle &= \sum_{n_1=0}^{n_0}\sum_{n_2=0}^{n_1}B_{n_0,p}(n_1)B_{n_1,p}(n_2)\frac{(n_1p-n_2)^2}{n_1}\\ &= \sum_{n_1=0}^{n_0}B_{n_0,p}(n_1)p(1-p)\\ &= p(1-p)\end{aligned}$$

This process can be continued iteratively to give the final result:

$$\begin{aligned}\langle S\rangle &= Mp(1-p)\\ \langle\nu_{\text{modal}}\rangle &= \nu_{\text{true}}\end{aligned} \tag{D.7}$$

Equation D.7 shows that if ν_{modal} is calculated according to equation D.5 and averaged over a sufficiently large number of experiments then the result will converge to the true value ν_{true}.

D.2 Bleaching Protocol

The bleaching protocol was performed following the standard lab protocol for a microscopy time lapse on slides, with some small caveats. Cells were fixed using the paraformaldehyde protocol outlined in appendix B, and adhered to slides using concanavalin A according to the standard lab protocol. To photobleach, the

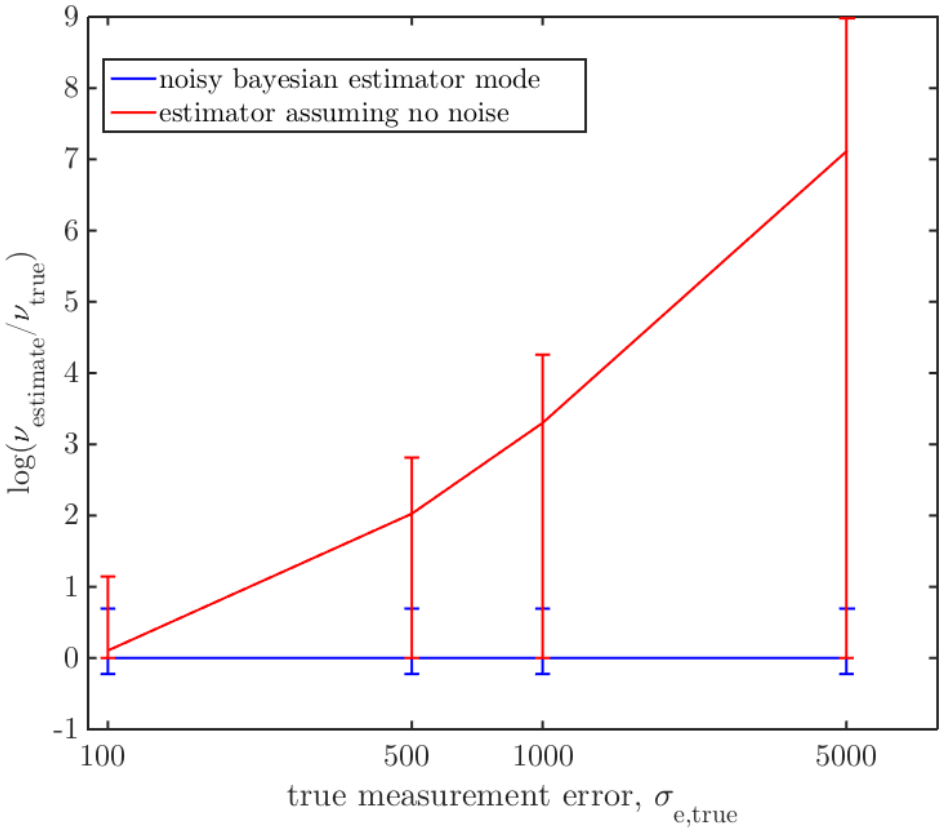

Figure D.1: Comparison of noiseless and noisy Bayesian estimator applied to simulated data with Gaussian noise. As described in the main text of chapter 4, sets of 30 cells were simulated with Gaussian measurement error of varying intensity. Both the noiseless analytic estimator given by equation D.5 and the Bayesian estimator described by equation 4.5 were applied and are shown here in red and blue respectively. For the Bayesian estimator the likelihood was calculated for a grid of values and the ν of the maximum a posteriori (ν, σ) grid point plotted. Since this was the same for multiple runs, the error is given as the width of the grid used at that grid point. It can be seen that while the Bayesian estimator performs well for all σ_e inspected, the noiseless analytic estimator dramatically over estimates ν for high $\sigma_{e,\text{true}}$. This is not suprising since a large σ_e will increase deviations from mean behaviour, which in the absence of noise would indicate low molecule numbers.

microscope software was to perform a acquire GFP (and GFPAutoFl if required) images every 10 seconds for twenty minutes at a single position. Brightfield was imaged only at the last time point. Acquisition was started and the GFP excitation LED immediately switched to 'on', bypassing the software and maintaining illumination throughout the time lapse.

After cells had been bleached, and empty area of the slide was found to acquire a background subtraction time lapse. This was performed in identical fashion to the that outlined above.

To maintain consistency, each slide was only used once, and if multiple acquisitions were acquired on the same day they were taken from separate slides.

Appendix E

Appendix to *GAL10*-lncRNA Investigation

E.1 Protocol for Induction Experiments with 3 Chamber ALCATRAS Device

The protocol is a slight modification of the standard 3 chamber protocol used in the lab. The 3 chamber device has only one inlet, making switching difficult. To observe induction, we loaded in the usual way, inserting cell syringes and pressuring them to hold cells agains the filters. When a sufficient number of cells were present, the medial inlet tub was removed and replaced with the one connected to the induction media. Cell syringes were depressurised and the syringe pumped. The acquisition was started as quickly as possible after this point, but we were not able to observe the arrival of inducing media.

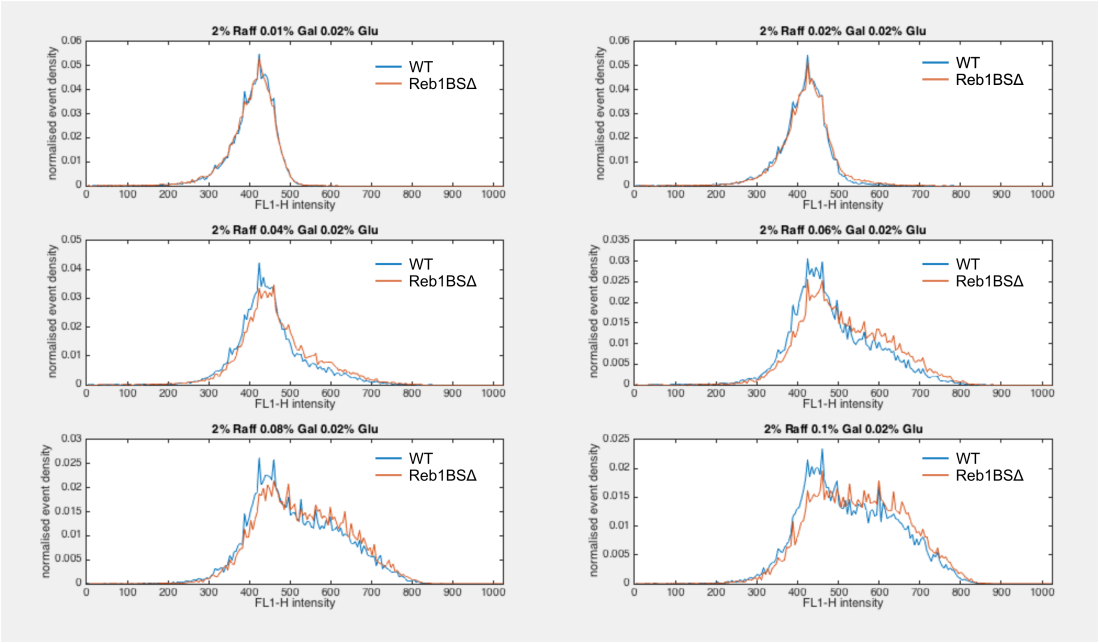

Figure E.1: In order to determine a regime in which cells both express robustly and show significant differences a range of sugar concentrations was investigated by flow cytometry. Cells were grown overnight in synthetic complete (SC) 2% raffinose, rediluted to an OD of 0.05, grown for a further 4 hours in SC 2% raffinose to an OD of approximately 0.2 and then centrifuged and resuspended in inducing media. In each case 2% raffinose and 0.02% glucose were used (as in Houseley et al. [84]) but the level of galactose was varied between the original concentration of 0.01% and 0.1%. Histograms of GFP fluorescence for the two strains in the various media is shown in the plots. 0.04% galactose concentration was chosen for further study since it was the concentration closest to the original which showed both robust expression and a significant difference between the two strains.

E.2 Selection of Appropriate Sugar Regime by Flow Cytometry

E.3 Autofluorescence Correction

To attempt to subtract autofluorescence, and as a precursor for identifying on and off cells, the data was corrected in two ways. First, to compensate for any differences in focus between the two chambers the data was normalised by the mean of the first 6 time points for each population. Since we expected no expres-

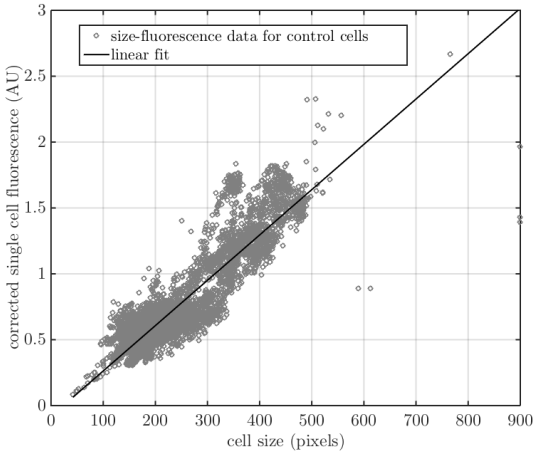

Figure E.2: Autofluorescence was estimated from cell size by a linear fit between control cell fluorescence and size. This estimated autofluorescence was subtracted from the measured value for each cell at each time point to give an estimated fluorescence due to GFP.

sion in this period, ww reasoned that all cells should be purely autofluorescent, and that any difference should be due to focus and but compensated by a multiplicative constant. The differences in these multiplicative constants was less than 2% of the total for the two fluorescent strains. Next, autofluorescence was estimated by fitting a straight line between size and fluorescence for all the data from the control cells. This linear fit was used to estimate the autofluorescence of WT and Reb1BSΔ cells using their size at each time point. This estimated autofluorescence was subtracted from the measured value to give a fluorescence due to GFP. Obviously this is a rather crude estimate, and sometimes produces negative values.

E.4 Processing of wild type* and Reb1BSΔ* Data

As in other experiments, cells were identified, segmented and data extracted by automated matlab routines. The sum of cellular pixels was used as an estimate of cellular fluorescence and only cells present for more than 200 timepoints (16

hours) were considered in the analysis. The multiplicative correction applied to wild type and Reb1BSΔ data was not felt to be appropriate because expression was so fast that cells could not be certainly said to be only autofluorescent during the first 6 time points. The same procedure (calculating three standard deviations of the control cells) was used for calculating a threshold for activation, but due to the pulsatile nature of the data the number of consecutive time points required for a cell to be considered ON was reduced to 5 (25 minutes).

For DPP analysis it was necessary to subtract autofluorescence. Reasoning that most cells would be autofluorescent at some point in the time lapse we took the minimum for each cell across the whole period of observation and subtracted this as an estimate of autofluorescence. To check for cells that fluorescently expressed throughout their time course we calculated the distribution of minima for control cells and any value falling outside two standard deviations of this was deemed untrustworthy. In these cases the mean of the control cell minima was subtracted from the cells trace instead. This autofluorescence corrected trace was divided by a protein:fluorescence ratio estimated from Hog1 data and whole proteome data sets [58]. The lognormal error model used by DPP has zero probability density for a measurement of 0, so fluorescence equivalent to 1 protein was added to the data to ensure positive values and prevent singularities.

E.5 details of DPP runs

Dynamic prior propagation (DPP) is an inference scheme developed by Zechner et al. to infer the parameters of models with extrinsically varying rate constants. In these models cellular processes are described as a standard chemical master equation (CME) but with each reaction rate being either intrinsic or extrinsic. Intrinsic reaction rates are common to all cells in the populations, whereas ex-

trinsic ones vary from cell to cell, are constant in time for each cell and are drawn for each cell from a hyper distribution with a set of hyper parameters.

DPP is a sequential monte carlo (SMC) [63] based inference algorithm. A large number of 'particles' are initialised to the initial state of the system (or random possible initial states of the system if it is unknown) - the system being all the cells observed. These particles are then resampled, using the probability of this initial state given the initial measurement as a weight by which to sample. Each particle is then simulated to the next time point using the gillespie algorithm and the process repeated. The liklihood of a particular set of parameters can be estimated by summing the weights before resampling and multiplying these weights over the whole timeseries.

In previous SMC schemes, this path sampling has been nested inside an MCMC scheme to sample parameters, so that the MCMC scheme samples parameters based on the likelihood estimate provided by the SMC scheme [63]. The strength of DPP is that it uses analytical results for the likelihood of both intrinsic and extrinisc parameters given the simulated system path to avoid the need for this MCMC wrapper. In this way no parameters are specified for any given particle. The final result is a likelihood of the intrinsic parameters and of the hyper parameters of the extrinsically varying reaction rates. Sample paths for the model species in the individual cells can also be generated, but this requires more particles to be used and is therefore more computationally intensive.

DPP was run largely according the default parameters. To number of particles was set to 10,000 , with 50% being discarded at each time point.

Since our condition were constant and our reporter shown to mature in under 5 minutes, no time delay to correct for protein folding was required.

Appendix F

Strains Used Throughout the Thesis

strain number	strain	purpose
188	trp1Δ :: Gal1pr(synthetic)/UBI-MΔk-GFP*	the synthetic Gal1/cyc5 promoter driving UBI-M-Δ k-GFP* from [85]. Intended to test fluorophore properties. Found to have very low protein expression.
189	trp1Δ :: Gal1pr(synthetic)/UBI-YΔk-GFP*	the synthetic Gal1/cyc5 promoter driving UBI-Y-Δ k-GFP* from [85]. Intended to test fluorophore properties. Found to have very low protein expression.

202	prs425[Gal1pr(synthetic)/UBI-MΔk-GFP*]	Cells harbouring the high copy number prs425 yeast plasmid containing the synthetic Gal1/cyc5 promoter driven UBI-M-Δ k-GFP* from [85] . Intended to test fluorophore properties. Found to have poor growth in galactose.
203	prs425[Gal1pr(synthetic)/UBI-YΔk-GFP*]	Cells harbouring the high copy number prs425 yeast plasmid containing the synthetic Gal1/cyc5 promoter driven UBI-Y-Δ k-GFP* from [85] . Intended to test fluorophore properties. Found to have poor growth in galactose.
223	(*gal1*Δ ::UBI-YΔk-GFP*)	expression of UBI-YΔkGFP* from endogenous *GAL1*pr
241	(*gal1*Δ ::UBI-MΔk-GFP*)	expression of UBI-MΔkGFP* from endogenous *GAL1*pr
222	(*gal1*Δ ::GFP*)	expression of GFP* from endogenous *GAL1*pr
91	*HOG1*-GFP	obtained from the GFP fusion collection and used in the bleaching project.
302	(*gal1*Δ ::UBI-MΔk-GFPγ)	expression of UBI-MΔkGFPγ from endogenous *GAL1*pr
304	(*gal1*Δ ::UBI-MΔkmKate2)	expression of UBI-MΔ-kmKate2 from endogenous *GAL1*pr
356	WT(tollervey lab) *GAL1*-EGFP	investigation of the effect of Gal10-lncRNA by comparison with strain 357
357	Reb1BSΔ *GAL1*-EGFP	investigation of the effect of Gal10-lncRNA by comparison with strain 356
384	WT(tollervey lab) *gal1*Δ ::UBI-MΔk-GFPγ	WT* in the main text. Used for investigation of the effect of *GAL*10-lncRNA by comparison with strain 385

385	Reb1BSΔ *gal1*Δ ::UBI-MΔk-GFPγ	Reb1BSΔ* in the main text. Used for investigation of the effect of *GAL*10-lncRNA by comparison with strain 384
386	WT(tollervey lab) *gal1*Δ ::UBI-MΔk-mKate2	control* in the main text. Used for investigation of the effect of *GAL*10-lncRNA by comparison with strain 384 and 385

Table F.1: table of strains constructed and obtained throughout the thesis.